# A Tensor Spectral Approach to Learning Mixed Membership Community Models

Anima Anandkumar[♯] Rong Ge[†] Daniel Hsu[‡] Sham M. Kakade[‡]
[♯]UC Irvine [†]Princeton University [‡]Microsoft Research

February 13, 2013

**Abstract**

Modeling community formation and detecting hidden communities in networks is a well studied problem. However, theoretical analysis of community detection has been mostly limited to models with non-overlapping communities such as the stochastic block model. In this paper, we remove this restriction, and consider a family of probabilistic network models with overlapping communities, termed as the mixed membership Dirichlet model, first introduced in [2]. This model allows for nodes to have fractional memberships in multiple communities and assumes that the community memberships are drawn from a Dirichlet distribution. We propose a unified approach to learning these models via a tensor spectral decomposition method. Our estimator is based on low-order moment tensor of the observed network, consisting of 3-star counts. Our learning method is fast and is based on simple linear algebra operations, e.g. singular value decomposition and tensor power iterations. We provide guaranteed recovery of community memberships and model parameters and present a careful finite sample analysis of our learning method. Additionally, our results match the best known scaling requirements in the special case of the stochastic block model.

**Keywords:** Community detection, spectral methods, tensor methods, moment-based estimation, mixed membership models.

## 1 Introduction

Studying communities forms an integral part of social network analysis. A community generally refers to a group of individuals with shared interests (e.g. music, sports), or relationships (e.g. friends, co-workers). Community formation in social networks has been studied by many sociologists, e.g. [33, 29, 31, 14], starting with the seminal work of Moreno [33]. They posit various factors such as *homophily*[1] among the individuals to be responsible for community formation. Various (probabilistic and non-probabilistic) network models attempt to explain community formation. In addition, they also attempt to quantify interactions and the extent of overlap between different communities, relative sizes among the communities, and various other network properties. Studying such community models are also of interest in other domains, e.g. in biological networks.

While there exists a vast literature on community models, learning these models is typically challenging, and various heuristics such as Markov Chain Monte Carlo (MCMC) or variational expectation maximization (EM) are employed in practice. Such heuristics tend to be unreliable and scale poorly for large networks. On the other hand, community models with guaranteed learning methods tend to be restrictive. A popular class of probabilistic models, termed as *stochastic blockmodels*, have been widely studied and enjoy strong theoretical learning guarantees, e.g. [42, 23, 16, 41, 37, 32]. On the other hand, they posit that an individual belongs to a single community, which does not hold in most real settings [34].

[1]The term *homophily* refers to the tendency that individuals belonging to the same community tend to connect more than individuals in different communities.

In this paper, we consider a class of mixed membership community models, originally introduced by Airoldi et. al. [2], and recently employed in [43, 21]. The model has been shown to be effective in many real-world settings, but so far, no learning approach exists with provable guarantees, and in practice, learning is carried out through Gibbs sampling or through variational Bayes. In this paper, we provide a novel learning approach for learning mixed membership models with provable guarantees.

The mixed membership community model of [2] has a number of attractive properties. It retains many of the convenient properties of the stochastic block model. For instance, conditional independence of the edges is assumed, given the community memberships of the nodes in the network. At the same time, it allows for communities to overlap, and for every individual to be fractionally involved in different communities. It includes the stochastic block model as a special case (corresponding to zero overlap among the different communities). This enables us to compare our learning guarantees with existing works for stochastic block models and also study how the extent of overlap among different communities affects the learning performance.

## 1.1 Summary of Results

We now summarize the main contributions of this paper. We propose a novel approach for learning mixed membership community models of [2, 43, 21]. Our approach is a method of moments estimator and incorporates tensor (spectral) decomposition. We provide guarantees for our approach under a set of sufficient conditions. Finally, we compare our results to existing ones for the special case of the stochastic block model, where nodes belong to a single community.

We present a unified approach for the mixed membership model of [2]. The extent of overlap between different communities in this model class is controlled (roughly) through a single scalar parameter, termed as the Dirichlet concentration parameter $\alpha_0 := \sum_i \alpha_i$, when the community vectors are drawn from the Dirichlet distribution $\text{Dir}(\alpha)$. When $\alpha_0 \to 0$, the mixed membership model degenerates to a stochastic block model. We propose a unified learning method and provide recovery guarantees under a set of sufficient conditions. We provide explicit scaling requirements in terms of the extent of community overlaps (through $\alpha_0$), the network size $n$, the number of communities $k$, and the average edge connectivity across various communities. For instance, for the special case, where $p$ is the probability for any intra-community edge to occur, and $q$ corresponds to the inter-community connectivity, and the average community sizes are equal, we require that[2]

$$n = \tilde{\Omega}(k^2(\alpha_0+1)^2), \qquad \frac{p-q}{\sqrt{p}} = \tilde{\Omega}\left(\frac{(\alpha_0+1)k}{n^{1/2}}\right). \tag{1}$$

Thus, we require $n$ to be large enough compared to the number of communities $k$, and for the separation $p-q$ to be large enough, so that the learning method can distinguish the different communities. Moreover, we see that the scaling requirements become more stringent as $\alpha_0$ increases. This is intuitive since it is harder to learn communities with more overlap, and we quantify this scaling. Moreover, we quantify the error bounds for estimating various parameters of the mixed membership model. Lastly, we establish zero-error guarantees for support recovery: our learning method correctly identifies (w.h.p) all the significant memberships of a node and also identifies the set of communities where a node does not have a strong presence, and we quantify these thresholds depending on $\alpha_0$.

For the special case of stochastic block models ($\alpha_0 \to 0$), (2) reduces to

$$n = \tilde{\Omega}(k^2), \qquad \frac{p-q}{\sqrt{p}} = \tilde{\Omega}\left(\frac{k}{n^{1/2}}\right), \tag{2}$$

The scaling requirements in (2) match with the best known bounds (up to poly-log factors) and were previously achieved by [44] via convex optimization involving semi-definite programming (SDP). In contrast, we propose an iterative non-convex approach involving tensor power iterations and linear algebraic techniques,

[2]The notation $\tilde{\Omega}(\cdot), \tilde{O}(\cdot)$ denotes $\Omega(\cdot), O(\cdot)$ up to log factors.

and obtain similar guarantees. For a detailed comparison of learning guarantees under various methods for learning stochastic block models, see [44].

Thus, we establish learning guarantees explicitly in terms of the extent of overlap among the different communities. Many real-world networks involve sparse community memberships and the total number of communities is typically much larger than the extent of membership of a single individual, e.g. hobbies/interests of a person, university/company networks that a person belongs to, the set of transcription factors regulating a gene, and so on. Thus, we see that in this regime of practical interest, where $\alpha_0 = \Theta(1)$, the scaling requirements in (1) match those for the stochastic block model in (2) (up to polylog factors) without any degradation in learning performance. Thus, we establish that learning community models with sparse community memberships is akin to learning stochastic block models and we present a unified approach and analysis for learning these models. To the best of our knowledge, this work is the first to establish polynomial time learning guarantees for probabilistic network models with overlapping communities and we provide a fast and an iterative learning approach through linear algebraic techniques and tensor power iterations.

## 1.2 Overview of Techniques

We now describe the main techniques employed in our learning approach and in establishing the recovery guarantees.

**Method of moments and subgraph counts:** We propose an efficient learning algorithm based on low order moments, viz., counts of small subgraphs. Specifically, we employ a third-order tensor which counts the number of 3-stars in the observed network. A 3-star is a star graph with three leaves (see figure 1) and we count the occurrences of such 3-stars across different partitions. We establish that (an adjusted) 3-star count tensor has a simple relationship with the model parameters, when the network is drawn from a mixed membership model. In particular, we propose a multi-linear transformation (also termed as whitening) under which the *canonical polyadic (CP) decomposition* of the tensor yields the model parameters and the community vectors. Note that the decomposition of a general tensor into its rank-one components is referred to as its CP decomposition [27] and is in general NP-hard [22]. However, we reduce the learning problem to an orthogonal symmetric tensor decomposition, for which tractable decomposition exists, as described below.

**Tensor spectral decomposition via power iterations:** Our tensor decomposition method is based on the popular power iterations (e.g. see [3]). It is a simple iterative method to compute the stable eigen-pairs of a tensor. In this paper, we propose various modifications to the basic power method to strengthen the recovery guarantees under perturbations. For instance, we introduce novel adaptive deflation techniques (which involves subtracting out the eigen-pairs which are previously estimated). Moreover, we optimize performance for the regime where the community overlaps are small. We initialize the tensor power method with (whitened) neighborhood vectors from the observed network. In the regime, where the community overlaps are small, this leads to an improved performance compared to random initialization. Additionally, we incorporate thresholding as a post-processing operation, which again, leads to improved guarantees for sparse community memberships, i.e., when the overlap among different communities is small.

**Sample analysis:** We establish that our learning approach correctly recovers the model parameters and the community memberships of all nodes under exact moments. We then carry out a careful analysis of the empirical graph moments, computed using the network observations. We establish tensor concentration bounds and also control the perturbation of the various quantities used by our learning algorithm via matrix Bernstein's inequality [40, thm. 1.4] and other inequalities. We impose the scaling requirements in (1) for various concentration bounds to hold.

## 1.3 Related Work

There is extensive work on modeling communities and various algorithms and heuristics for discovering them. We mostly limit our focus to works with theoretical guarantees.

**Method of moments:** The method of moments approach dates back to Pearson [35] and has been applied for learning various community models. Here, the moments correspond to counts of various subgraphs in the network. They typically consist of aggregate quantities, e.g., number of star subgraphs, triangles etc. in the network. For instance, Bickel *et al* [9] analyze the moments of a stochastic block model and establish that the subgraph counts of certain structures, termed as "wheels" (a family of trees), are sufficient for identifiability under some natural non-degeneracy conditions. In contrast, we establish that moments up to third order (corresponding to edge and 3-star counts) are sufficient for identifiability of the stochastic block model, and also more generally, for the mixed membership Dirichlet model. We employ subgraph count tensors, corresponding to the number of subgraphs (such as stars) over a set of labeled vertices, while the work in [9] considers only aggregate (i.e. scalar) counts. Considering tensor moments allows us to use simple subgraphs (edges and 3 stars) corresponding to low order moments, rather than more complicated graphs (e.g. wheels considered in [9]) with larger number of nodes, for learning the community model.

The method of moments is also relevant in the context of a family of random graph models termed as *exponential random graph models* [24, 17]. Subgraph counts of fixed graphs such as stars and triangles serve as sufficient statistics for these models. However, parameter estimation given the subgraph counts is in general NP-hard, due to the normalization constant in the likelihood (the partition function) and the model suffers from degeneracy issues; see [36, 13] for detailed discussion. In contrast, we establish in this paper that the mixed membership model is amenable to simple estimation methods through linear algebraic operations and tensor power iterations using subgraph counts of 3-stars.

**Stochastic block models:** Many algorithms provide learning guarantees for stochastic block models. For a detailed comparison of these methods, see the recent work in [44]. A popular method is based on spectral clustering [32], where community memberships are inferred through projection onto the spectrum of the Laplacian matrix (or its variants). This method is fast and easy to implement (via singular value decomposition). On the other hand, its theoretical learning guarantees are worse compared to the work of [44], which uses convex optimization techniques via semi-definite programming. For a detailed comparison of learning guarantees under various methods for learning stochastic block models, see [44].

**Non-probabilistic approaches:** The classical approach to community detection tries to directly exploit the properties of the graph to define communities, without assuming a probabilistic model. Girvan and Newman [20] use betweenness to remove edges until only communities are left. However, Bickel and Chen [8] show that these algorithms are (asymptotically) biased and that using modularity scores can lead to the discovery of an incorrect community structure, even for large graphs. Jalali *et al* [25] define community structure as the structure that satisfies the maximum number of edge constraints (whether two individuals like/dislike each other). However, these models assume that every individual belongs to a single community.

Recently, some non-probabilistic approaches have been introduced with overlapping community models by Arora *et al* [6] and Balcan *et al* [7]. The analysis of Arora *et al* [6] is mostly limited to dense graphs (i.e. $\Theta(n^2)$ edges for a $n$ node graph), while our analysis provides learning guarantees for much sparser graphs (as seen by the scaling requirements in (1)). Moreover, the running time of the method in [6] is *quasipolynomial time* (i.e. $O(n^{\log n})$) for the general case, and is based on a combinatorial learning approach. In contrast, our learning approach is based on simple linear algebraic techniques and the running time is a low-order polynomial (roughly it is $O(n^2 k)$ for a $n$ node network with $k$ communities). The work of [7] assumes endogenous formed communities, by constraining the fraction of edges within a community compared to the outside. They provide a polynomial time algorithm for finding all such "self-determined" communities and the running time is $n^{O(\log 1/\alpha)/\alpha}$, where $\alpha$ is the fraction of edges within a self-determined community, and this bound is improved to linear time when $\alpha > 1/2$. On the other hand, the running time of our algorithm is

mostly independent of the parameters of the assumed model, (and is roughly $O(n^2k)$). Moreover, both these works are limited to homophilic models, where there are more edges within each community, than between any two different communities. However, our learning approach is not limited to this setting and also does not assume homogeneity in edge connectivity across different communities (while indeed it makes probabilistic assumptions on community formation). In addition, we provide improved guarantees for homophilic models by considering additional post-processing steps in our algorithm. Recently, Abraham *et al* [1] provide an algorithm for approximating the parameters of an Euclidean log-linear model in polynomial time. However, there setting is considerably different than the one in this paper.

**Inhomogeneous random graphs, graph limits and weak regularity lemma:** Inhomogeneous random graphs have been analyzed in a variety of settings (e.g., [11, 30]) and are generalizations of the stochastic block model. Here, the probability of an edge between any two nodes is characterized by a general function (rather than by a $k \times k$ matrix as in the stochastic block model with $k$ blocks). Note that the mixed membership model considered in this work is a special instance of this general framework. These models arise as the limits of convergent (dense) graph sequences and for this reason, the functions are also termed as "graphons" or graph limits [30]. A deep result in this context is the regularity lemma and its variants. The weak regularity lemma proposed by Frieze and Kannan [18], showed that any convergent dense graph can be approximated by a stochastic block model. Moreover, they propose an algorithm to learn such a block model based on the so-called $d_2$ distance. The $d_2$ distance between two nodes measures similarity with respect to their "two-hop" neighbors and the block model is obtained by thresholding the $d_2$ distances. However, the method is limited to learning block models and not overlapping communities.

**Learning Latent Variable Models (Topic Models):** We leverage the recent developments from [5, 3, 4] for learning topic models and other latent variable models based on the method of moments. Topic models have been popular in document modeling [10], and they posit that the words in a document are generated through multiple latent topics in a document. The works in [5, 3, 4] consider learning these models from second- and third-order observed moments through linear algebraic and tensor-based techniques. We exploit the tensor power iteration method of [4] and provide additional improvements to obtain stronger recovery guarantees. Moreover, the sample analysis is quite different in the community setting compared to other latent variable models analyzed in [5, 3, 4].

**Relationship between community detection and tensor decomposition:** There are works relating the hardness of finding hidden cliques and the use of higher order moment tensors for this purpose. Frieze and Kannan [19] relate the problem of finding a hidden clique to finding the top eigenvector of the third order tensor, corresponding to the maximum spectral norm. However, this problem is known to be NP-hard in general [22]. Brubaker and Vempala [12] extend the result to arbitrary $r^{\text{th}}$-order tensors. The work in [15] provides lower bounds on the complexity of statistical algorithms, i.e., those with access to moments (rather than actual samples from the distribution) and show that the cliques have to be size $\Omega(n^{1/r})$ to enable recovery from $r^{\text{th}}$-order moment tensors in a $n$ node network. Thus tensors are useful for finding smaller hidden cliques in network (albeit by solving a computationally hard problem in general). In contrast, we consider tractable tensor decomposition through reduction to orthogonal symmetric tensors (under the scaling requirements of (1)), and our learning method is a fast and an iterative approach based on tensor power iterations and linear algebraic operations.

# 2 Community Models and Graph Moments

## 2.1 Community Membership Models

In this section, we describe the mixed membership community model based on Dirichlet priors for the community draws by the individuals. We first introduce the special case of the popular stochastic block model, where each node belongs to a single community.

**Notation:** We consider stochastic network models with $n$ nodes and let $[n] := \{1, 2, \ldots, n\}$. Let $G$ be the $\{0, 1\}$ adjacency[3] matrix for the random network and let $G_{A,B}$ be the submatrix of $G$ corresponding to rows $A \subseteq [n]$ and columns $B \subseteq [n]$. We consider models with $k$ underlying (hidden) communities. For node $i$, let $\pi_i \in \mathbb{R}^k$ denote its *community membership vector*, i.e., the vector is supported on the communities to which the node belongs. In the special case of the popular stochastic block model described below, $\pi_i$ is a basis coordinate vector, while the more general mixed membership model relaxes this assumption and a node can be in multiple communities with fractional memberships. Define $\Pi := [\pi_1|\pi_2|\cdots|\pi_n] \in \mathbb{R}^{k\times n}$. and let $\Pi_A := [\pi_i : i \in A] \in \mathbb{R}^{k\times|A|}$ denote the set of column vectors restricted to $A \subseteq [n]$. For a matrix $M$, let $(M)_i$ and $(M)^i$ denote its $i^{\text{th}}$ column and row respectively. For a matrix $M$ with singular value decomposition (SVD) $M = UDV^\top$, let $(M)_{k-svd} := U\tilde{D}V^\top$ denote the $k$-rank SVD of $M$, where $\tilde{D}$ is limited to top-$k$ singular values of $M$. Let $M^\dagger$ denote the MoorePenrose pseudo-inverse of $M$. Let $\mathbb{I}(\cdot)$ be the indicator function. We use the term high probability to mean with probability $1 - k^{-c}$ for any constant $c > 0$.

**Stochastic block model (special case):** In this model, each individual is independently assigned to a single community, chosen at random: each node $i$ chooses community $j$ independently with probability $\widehat{\alpha}_j$, for $i \in [n], j \in [k]$, and we assign $\pi_i = e_j$ in this case, where $e_j \in \{0, 1\}^k$ is the $j^{\text{th}}$ coordinate basis vector. Given the community assignments $\Pi$, every directed[4] edge in the network is independently drawn: if node $u$ is in community $i$ and node $v$ is in community $j$ (and $u \neq v$), then the probability of having the edge $(u, v)$ in the network is $P_{i,j}$. Here, $P \in [0, 1]^{k\times k}$ and we refer to it as the *community connectivity matrix*. This implies that given the community membership vectors $\pi_u$ and $\pi_v$, the probability of an edge from $u$ to $v$ is $\pi_u^\top P\pi_v$ (since when $\pi_u = e_i$ and $\pi_v = e_j$, we have $\pi_u^\top P\pi_v = P_{i,j}$.). The stochastic model has been extensively studied and can be learnt efficiently through various methods, e.g. spectral clustering [32], convex optimization [44]. and so on. Many of these methods rely on conditional independence assumptions of the edges in the block model for guaranteed learning.

**Mixed membership model:** We now consider the extension of the stochastic block model which allows for an individual to belong to multiple communities and yet preserves some of the convenient independence assumptions of the block model. In this model, the community membership vector $\pi_u$ at node $u$ is a probability vector, i.e., $\sum_{i\in[k]} \pi_u(i) = 1$, for all $u \in [n]$. Given the community membership vectors, the generation of the edges is identical to the block model: given vectors $\pi_u$ and $\pi_v$, the probability of an edge from $u$ to $v$ is $\pi_u^\top P\pi_v$, and the edges are independently drawn. Under this formulation, given the community vectors $\Pi$, a random network can be generated as follows: for each node pair $u, v$, a community pair $i, j$ is drawn independently from their community membership vectors, i.e. $i \sim \pi_u$ and $j \sim \pi_v$, and the edge $(u, v)$ is drawn independently with probability $P_{i,j}$. This formulation allows for the nodes to be in multiple communities, and at the same time, preserves the conditional independence of the edges, given the community memberships of the nodes.

**Dirichlet prior for community membership:** The only aspect left to be specified for the mixed membership model is the distribution from which the community membership vectors $\Pi$ are drawn. We consider the popular setting of [2, 21], where the community vectors $\{\pi_u\}$ are i.i.d. draws from the Dirichlet distribution, denoted by $\mathrm{Dir}(\alpha)$, with parameter vector $\alpha \in \mathbb{R}^k_{>0}$. The probability density function of the Dirichlet distribution is given by

$$\mathbb{P}[\pi] = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\alpha_0)} \prod_{i=1}^k \pi_i^{\alpha_i - 1}, \quad \pi \sim \mathrm{Dir}(\alpha), \alpha_0 := \sum_i \alpha_i, \tag{3}$$

where $\Gamma(\cdot)$ is the Gamma function and the ratio of the Gamma functions serves as the normalization constant.

[3]Our analysis can easily be extended to weighted adjacency matrices with bounded entries.

[4]We limit our discussion to directed networks in this paper, but note that the results also hold for undirected community models, where $P$ is a symmetric matrix, and an edge $(u, v)$ is formed with probability $\pi_u^\top P\pi_v = \pi_v^\top P\pi_u$.

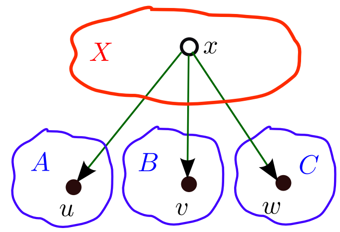


Figure 1: Our moment-based learning algorithm uses 3-star count tensor from partition $X$ to partitions $A, B, C$ (and the roles of the partitions are interchanged to get various estimates). Specifically, T is a third order tensor, where $\mathrm{T}(u, v, w)$ is the normalized count of the 3-stars with $u, v, w$ as leaves over all $x \in X$.

The Dirichlet distribution is widely employed for specifying priors in Bayesian statistics, e.g. latent Dirichlet allocation [10]. The Dirichlet distribution is the conjugate prior of the multinomial distribution which makes it attractive for Bayesian inference.

Let $\widehat{\alpha}$ denote the normalized parameter vector $\alpha/\alpha_0$, where $\alpha_0 := \sum_i \alpha_i$. In particular, note that $\widehat{\alpha}$ is a probability vector: $\sum_i \widehat{\alpha}_i = 1$. Intuitively, $\widehat{\alpha}$ denotes the relative expected sizes of the communities (since $\mathbb{E}[n^{-1} \sum_{u \in [n]} \pi_u[i]] = \widehat{\alpha}_i$). Let $\widehat{\alpha}_{\max}$ be the largest entry in $\widehat{\alpha}$, and $\widehat{\alpha}_{\min}$ be the smallest entry. Our learning guarantees will depend on these parameters.

The stochastic block model is a limiting case of the mixed membership model when the Dirichlet parameter is $\alpha = \alpha_0 \cdot \widehat{\alpha}$, where the probability vector $\widehat{\alpha}$ is held fixed and $\alpha_0 \to 0$. In the other extreme when $\alpha_0 \to \infty$, the Dirichlet distribution becomes peaked around a single point, for instance if $\alpha_i \equiv c$ and $c \to \infty$, the Dirichlet distribution is peaked at $k^{-1} \cdot \vec{1}$, where $\vec{1}$ is the all-ones vector. Thus, the parameter $\alpha_0$ serves as a measure of the average sparsity of the Dirichlet draws or equivalently, of how concentrated the Dirichlet measure is along the different coordinates. This in effect, controls the extent of overlap among different communities.

**Sparse regime of Dirichlet distribution:** When the Dirichlet parameter vector satisfies[5] $\alpha_i < 1$, for all $i \in [k]$, the Dirichlet distribution $\mathrm{Dir}(\alpha)$ generates "sparse" vectors with high probability[6]; see [39] (and in the extreme case of the block model where $\alpha_0 \to 0$, it generates 1-sparse vectors). Many real-world settings involve sparse community membership and the total number of communities is typically much larger than the extent of membership of a single individual, e.g. hobbies/interests of a person, university/company networks that a person belongs to, the set of transcription factors regulating a gene, and so on. Our learning guarantees are limited to the sparse regime of the Dirichlet model.

## 2.2 Graph Moments Under Mixed Membership Models

Our approach for learning a mixed membership community model relies on the form of the graph moments[7] under the mixed membership model. We now describe the specific graph moments used by our learning algorithm (based on 3-star and edge counts) and provide explicit forms for the moments, assuming draws from a mixed membership model.

### Notations

Recall that $G$ denotes the adjacency matrix and that $G_{X,A}$ denotes the submatrix corresponding to edges going from $X$ to $A$. Recall that $P \in [0,1]^{k \times k}$ denotes the community connectivity matrix. Define

$$F := \Pi^\top P^\top = [\pi_1 | \pi_2 | \cdots | \pi_n]^\top P^\top. \tag{4}$$

[5] The assumption that the Dirichlet distribution be in the sparse regime is not strictly needed. Our results can be extended to general Dirichlet distributions, but with worse scaling requirements on the network size $n$ for guaranteed learning.

[6] Roughly the number of entries in $\pi$ exceeding a threshold $\tau$ is at most $O(\alpha_0 \log(1/\tau))$ with high probability, when $\pi \sim \mathrm{Dir}(\alpha)$.

[7] We interchangeably use the term first order moments for edge counts and third order moments for 3-star counts.

For a subset $A \subseteq [n]$ of individuals, let $F_A \in \mathbb{R}^{|A|\times k}$ denote the submatrix of $F$ corresponding to nodes in $A$, *i.e.*, $F_A := \Pi_A^\top P^\top$. Let $\mathrm{Diag}(v)$ denote a diagonal matrix with diagonal entries given by a vector $v$.

Our learning algorithm uses moments up to the third-order, represented as a tensor. A third-order tensor $T$ is a three-dimensional array whose $(p,q,r)$-th entry denoted by $T_{p,q,r}$. The symbol $\otimes$ denotes the standard Kronecker product: if $u$, $v$, $w$ are three vectors, then

$$(u \otimes v \otimes w)_{p,q,r} := u_p \cdot v_q \cdot w_r. \tag{5}$$

A tensor of the form $u \otimes v \otimes w$ is referred to as a rank-one tensor. The decomposition of a general tensor into a sum of its rank-one components is referred to as *canonical polyadic (CP) decomposition* [27]. We will subsequently see that the graph moments can be expressed as a tensor and that the CP decomposition of the graph-moment tensor yields the model parameters and the community vectors under the mixed membership community model.

#### 2.2.1 Graph moments under Stochastic Block Model

We first analyze the graph moments in the special case of a stochastic block model (i.e., $\alpha_0 = \sum_i \alpha_i \to 0$ in the Dirichlet prior in (3)) and then extend it to general mixed membership model. We provide explicit expressions for the graph moments corresponding to edge counts and 3-star counts. We later establish in Section 3 that these moments are sufficient to learn the community memberships of the nodes and the model parameters of the block model.

**3-star counts:** The primary quantity of interest is a third-order tensor which counts the number of 3-stars. A 3-star is a star graph with three leaves $\{a,b,c\}$ and we refer to the internal node $x$ of the star as its "head", and denote the structure by $x \to \{a,b,c\}$ (see figure 1). We partition the network into four[8] parts and consider 3-stars such that each node in the 3-star belongs to a different partition. This is necessary to obtain a simple form of the moments, based on the conditional independence assumptions of the block model, see Proposition 2.1. Specifically, consider a partition $A, B, C, X$ of the network. We count the number of 3-stars from $X$ to $A, B, C$ and our quantity of interest is

$$\mathrm{T}_{X\to\{A,B,C\}} := \frac{1}{|X|}\sum_{i\in X}[G_{i,A}^\top \otimes G_{i,B}^\top \otimes G_{i,C}^\top], \tag{6}$$

where $\otimes$ is the Kronecker product, defined in (5) and $G_{i,A}$ is the row vector supported on the set of neighbors of $i$ belonging to set $A$. $\mathrm{T} \in \mathbb{R}^{|A|\times|B|\times|C|}$ is a third order tensor, and an element of the tensor is given by

$$\mathrm{T}_{X\to\{A,B,C\}}(a,b,c) = \frac{1}{|X|}\sum_{x\in X} G(x,a)G(x,b)G(x,c), \quad \forall a\in A, b\in B, c\in C, \tag{7}$$

which is the normalized count of the number of 3-stars with leaves $a, b, c$ such that its "head" is in set $X$.

We now relate the tensor $\mathrm{T}_{X\to\{A,B,C\}}$ to the parameters of the stochastic block model, viz., the community connectivity matrix $P$ and the community probability vector $\widehat{\alpha}$, where $\widehat{\alpha}_i$ is the probability of choosing community $i$.

**Proposition 2.1** (Moments in Stochastic Block Model)**.** *Given partitions $A, B, C, X$, and $F := \Pi^\top P^\top$, where $P$ is the community connectivity matrix and $\Pi$ is the matrix of community membership vectors, we have*

$$\mathbb{E}[G_{X,A}^\top|\Pi_A, \Pi_X] = F_A \Pi_X, \tag{8}$$

$$\mathbb{E}[\mathrm{T}_{X\to\{A,B,C\}}|\Pi_A,\Pi_B,\Pi_C] = \sum_{i\in[k]} \widehat{\alpha}_i (F_A)_i \otimes (F_B)_i \otimes (F_C)_i, \tag{9}$$

*where $\widehat{\alpha}_i$ is the probability for a node to select community $i$.*

[8]For sample complexity analysis, we require dividing the graph into more than four partitions to deal with statistical dependency issues, and we outline it in Section 3.

**Remark 1:** Note the form of the 3-star count tensor T in (9). It provides a CP decomposition of T since each term in the summation, viz., $\widehat{\alpha}_i(F_A)_i \otimes (F_B)_i \otimes (F_C)_i$, is a rank one tensor. Thus, we can learn the matrices $F_A, F_B, F_C$ and the vector $\widehat{\alpha}$ through CP decomposition of tensor T. We can then exploit the form of the adjacency submatrix $G_{X,A}$ in (8) to obtain $\Pi_X$, the set of community vectors of nodes in $X$. Similarly, we can consider another tensor consisting of 3-stars from $A$ to $X, B, C$, and obtain matrices $F_X, F_B$ and $F_C$ through a CP decomposition, and so on. Once we obtain matrices $F$ and $\Pi$ for the entire set of nodes in this manner, we can obtain the community connectivity matrix $P$, since $F := \Pi^\top P^\top$. Thus, in principle, we are able to learn all the model parameters ($\widehat{\alpha}$ and $P$) and the community membership matrix $\Pi$ under the stochastic block model. This forms our basic approach for learning the community model using the adjacency matrix $G$ and the 3-star count tensor T. The details are in Section 3.

**Remark 2:** The main property exploited in proving the tensor form in (9) is the conditional-independence assumption under the stochastic block model: the realization of the edges in each 3-star, say in $x \to \{a, b, c\}$, is conditionally independent given the community membership vector $\pi_x$, when $x \neq a \neq b \neq c$. This is because the community membership vectors $\Pi$ are assumed to be drawn independently at the different nodes and the edges are drawn independently given the community vectors. Considering 3-stars from $X$ to $A, B, C$ where $X, A, B, C$ form a partition ensures that this conditional independence is satisfied for all the 3-stars in tensor T.

*Proof:* Recall that the probability of an edge from $u$ to $v$ given $\pi_u, \pi_v$ is

$$\mathbb{E}[G_{u,v}|\pi_u, \pi_v] = \pi_u^\top P \pi_v = \pi_v^\top P^\top \pi_u = F_v \pi_u,$$

and $\mathbb{E}[G_{X,A}|\Pi_A, \Pi_X] = \Pi_X^\top P \Pi_A = \Pi_X^\top F_A^\top$ and thus (8) holds. For the tensor form, first consider an element of the tensor, with $a \in A, b \in B, c \in C$,

$$\mathbb{E}\left[\mathrm{T}_{X\to\{A,B,C\}}(a, b, c)|\pi_a, \pi_b, \pi_c, \pi_x\right] = \frac{1}{|X|}\sum_{x\in X} F_a\pi_x \cdot F_b\pi_x \cdot F_c\pi_x,$$

where (a) follows from the conditional-independence assumption of the edges (assuming $a \neq b \neq c$). Now taking expectation over the nodes in $X$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{T}_{X\to\{A,B,C\}}(a, b, c)|\pi_a, \pi_b, \pi_c\right] &= \frac{1}{|X|}\sum_{x\in X} \mathbb{E}[F_a\pi_x \cdot F_b\pi_x \cdot F_c\pi_x] \\
&= \mathbb{E}[F_a\pi \cdot F_b\pi \cdot F_c\pi] \\
&= \sum_{j\in[k]} \widehat{\alpha}_j (F_a)_j \cdot (F_b)_j \cdot (F_c)_j,
\end{aligned}$$

where the last step follows from the fact that $\pi = e_j$ with probability $\widehat{\alpha}_j$ and the result holds when $x \neq a, b, c$. Recall that $(F_a)_j$ denotes the $j^{\text{th}}$ column of $F_a$ (since $F_a e_j = (F_a)_j$). Collecting all the elements of the tensor, we obtain the desired result. □

#### 2.2.2 Graph Moments under Mixed Membership Dirichlet Model

We now analyze the graph moments for the general mixed membership Dirichlet model. Instead of the raw moments (i.e. edge and 3-star counts), we consider modified moments to obtain similar expressions as in the case of the stochastic block model.

Let $\mu_{X\to A} \in \mathbb{R}^{|A|}$ denote a vector which gives the normalized count of edges from $X$ to $A$:

$$\mu_{X\to A} := \frac{1}{|X|}\sum_{i\in X}[G_{i,A}^\top]. \tag{10}$$

We now define a modified adjacency matrix[9] $G^{\alpha_0}_{X,A}$ as

$$G^{\alpha_0}_{X,A} := \left(\sqrt{\alpha_0+1}G_{X,A} - (\sqrt{\alpha_0+1}-1)\vec{1}\mu^\top_{X\to A}\right). \tag{11}$$

In the special case of the stochastic block model ($\alpha_0 \to 0$), $G^{\alpha_0}_{X,A} = G_{X,A}$ is the submatrix of the adjacency matrix $G$. Similarly, we define modified third-order statistics,

$$\begin{aligned}\mathrm{T}^{\alpha_0}_{X\to\{A,B,C\}} :=& (\alpha_0+1)(\alpha_0+2)\,\mathrm{T}_{X\to\{A,B,C\}} + 2\,\alpha_0^2\,\mu_{X\to A}\otimes\mu_{X\to B}\otimes\mu_{X\to C} \\ &- \frac{\alpha_0(\alpha_0+1)}{|X|}\sum_{i\in X}\left[G^\top_{i,A}\otimes G^\top_{i,B}\otimes\mu_{X\to C} + G^\top_{i,A}\otimes\mu_{X\to B}\otimes G^\top_{i,C} + \mu_{X\to A}\otimes G^\top_{i,B}\otimes G^\top_{i,C}\right],\end{aligned} \tag{12}$$

and it reduces to the 3-star count $\mathrm{T}_{X\to\{A,B,C\}}$ defined in (6) for the stochastic block model ($\alpha_0 \to 0$). The modified adjacency matrix and the 3-star count tensor can be viewed as a form of "centering" of the raw moments which simplifies the expressions for the moments. The following relationships hold between the modified graph moments $G^{\alpha_0}_{X,A}$, $\mathrm{T}^{\alpha_0}$ and the model parameters $P$ and $\widehat{\alpha}$ of the mixed membership model.

**Proposition 2.2** (Moments in Mixed Membership Model)**.** *Given partitions $A, B, C, X$ and $G^{\alpha_0}_{X,A}$ and $\mathrm{T}^{\alpha_0}$, as in* (11) *and* (12)*, normalized Dirichlet vector $\widehat{\alpha}$, and $F := \Pi^\top P^\top$, where $P$ is the community connectivity matrix and $\Pi$ is the matrix of community membership vectors, we have*

$$\mathbb{E}[(G^{\alpha_0}_{X,A})^\top|\Pi_A,\Pi_X] = F_A\,\mathrm{Diag}(\widehat{\alpha}^{1/2})\Psi_X, \tag{13}$$

$$\mathbb{E}[\mathrm{T}^{\alpha_0}_{X\to\{A,B,C\}}|\Pi_A,\Pi_B,\Pi_C] = \sum_{i=1}^k \widehat{\alpha}_i (F_A)_i\otimes(F_B)_i\otimes(F_C)_i, \tag{14}$$

*where $(F_A)_i$ corresponds to $i^{th}$ column of $F_A$ and $\Psi_X$ relates to the community membership matrix $\Pi_X$ as*

$$\Psi_X := \mathrm{Diag}(\widehat{\alpha}^{-1/2})\left(\sqrt{\alpha_0+1}\Pi_X - (\sqrt{\alpha_0+1}-1)\left(\frac{1}{|X|}\sum_{i\in X}\pi_i\right)\vec{1}^\top\right).$$

*Moreover, we have that*

$$|X|^{-1}\mathbb{E}_{\Pi_X}[\Psi_X\Psi_X^\top] = I. \tag{15}$$

**Remark:** The 3-star count tensor $T^{\alpha_0}$ is carefully chosen so that the CP decomposition of the tensor directly yields the matrices $F_A, F_B, F_C$ and $\widehat{\alpha}_i$, as in the case of the stochastic block model. Similarly, the modified adjacency matrix $(G^{\alpha_0}_{X,A})^\top$ is carefully chosen to eliminate second-order correlation in the Dirichlet distribution and we have that $|X|^{-1}\mathbb{E}_{\Pi_X}[\Psi\Psi^\top] = I$ is the identity matrix. These properties will be exploited by our learning algorithm in Section 3.

*Proof:* The proof is on lines of Proposition 2.1 for stochastic block models ($\alpha_0 \to 0$) but more involved due to the form of Dirichlet moments. Recall $\mathbb{E}[G^\top_{i,A}|\pi_i,\Pi_A] = F_A\pi_i$ for a mixed membership model, and $\mu_{X\to A} := \frac{1}{|X|}\sum_{i\in X}G^\top_{i,A}$, therefore $\mathbb{E}[\mu_{X\to A}|\Pi_A,\Pi_X] = F_A\left(\frac{1}{|X|}\sum_{i\in X}\pi_i\right)\vec{1}^\top$. Equation (13) follows directly. For Equation (15), we note the Dirichlet moment, $\mathbb{E}[\pi\pi^\top] = \frac{1}{\alpha_0+1}\mathrm{Diag}(\widehat{\alpha}) + \frac{\alpha_0}{\alpha_0+1}\widehat{\alpha}\widehat{\alpha}^\top$, when $\pi\sim\mathrm{Dir}(\alpha)$ and

$$\begin{aligned}|X|^{-1}\mathbb{E}[\Psi_X\Psi_X^\top] &= \mathrm{Diag}(\widehat{\alpha}^{-1/2})\left[(\alpha_0+1)\mathbb{E}[\pi\pi^\top] + (-2\sqrt{\alpha_0+1}(\sqrt{\alpha_0+1}-1)\right.\\ &\quad \left.+(\sqrt{\alpha_0+1}-1)^2)\mathbb{E}[\pi]\mathbb{E}[\pi]^\top\right]\mathrm{Diag}(\widehat{\alpha}^{-1/2}) \\ &= \mathrm{Diag}(\widehat{\alpha}^{-1/2})\left(\mathrm{Diag}(\widehat{\alpha}) + \alpha_0\widehat{\alpha}\widehat{\alpha}^\top + (-\alpha_0)\widehat{\alpha}\widehat{\alpha}^\top\right)\mathrm{Diag}(\widehat{\alpha}^{-1/2}) \\ &= I.\end{aligned}$$

[9] To compute the modified moments $G^{\alpha_0}$, and $\mathrm{T}^{\alpha_0}$, we need to know the value of the scalar $\alpha_0 := \sum_i \alpha_i$, which is the concentration parameter of the Dirichlet distribution and is a measure of the extent of overlap between the communities. We assume its knowledge here.

On lines of the proof of Proposition 2.1 for the block model, the expectation in (14) involves multi-linear map of the expectation of the tensor products $\pi \otimes \pi \otimes \pi$ among other terms. Collecting these terms, we have that

$$(\alpha_0+1)(\alpha_0+2)\mathbb{E}[\pi\otimes\pi\otimes\pi]-(\alpha_0)(\alpha_0+1)(\mathbb{E}[\pi\otimes\pi\otimes\mathbb{E}[\pi]]+\mathbb{E}[\pi\otimes\mathbb{E}[\pi]\otimes\pi]+\mathbb{E}[\mathbb{E}[\pi]\otimes\pi\otimes\pi])+2\alpha_0^2\mathbb{E}[\pi]\otimes\mathbb{E}[\pi]\otimes\mathbb{E}[\pi]$$

is a diagonal tensor, in the sense that its $(p,p,p)$-th entry is $\widehat{\alpha}_p$, and its $(p,q,r)$-th entry is 0 when $p,q,r$ are not all equal. With this, we have (14). □

Note the nearly identical forms of the graph moments for the stochastic block model in (8), (9) and for the general mixed membership model in (13), (14). In other words, the modified moments $G^{\alpha_0}_{X,A}$ and $\mathrm{T}^{\alpha_0}$ have similar relationships to underlying parameters as the raw moments in the case of the stochastic block model. This enables us to use a unified learning approach for the two models, outlined in the next section.

# 3 Algorithm for Learning Mixed Membership Models

The simple form of the graph moments derived in the previous section is now utilized to recover the community vectors $\Pi$ and model parameters $P, \widehat{\alpha}$ of the mixed membership model. The method is based on the so-called tensor power method, used to obtain a tensor decomposition. We first outline the basic tensor decomposition method below and then demonstrate how the method can be adapted to learning using the graph moments at hand. We first analyze the simpler case when exact moments are available in Section 3.2 and then extend the method to handle empirical moments computed from the network observations in Section 3.3.

## 3.1 Overview of Tensor Decomposition Through Power Iterations

In this section, we review the basic method for tensor decomposition based on power iterations for a special class of tensors, viz., symmetric orthogonal tensors. Subsequently, in Section 3.2 and 3.3, we modify this method to learn the mixed membership model from graph moments, described in the previous section. For details on the tensor power method described below, refer to [3, 28].

Recall that a third-order tensor $T$ is a three-dimensional array and we use $T_{p,q,r}$ to denote the $(p,q,r)$-th entry of the tensor $T$. The standard symbol $\otimes$ is used to denote the Kronecker product, and $(u \otimes v \otimes w)$ is a rank one tensor. The decomposition of a tensor into its rank one components is called the CP decomposition.

**Multi-linear maps:** We can view a tensor $T \in \mathbb{R}^{d\times d\times d}$ as a multilinear map in the following sense: for a set of matrices $\{V_i \in \mathbb{R}^{d\times m_i} : i \in [3]\}$, the $(i_1,i_2,i_3)$-th entry in the three-way array representation of $T(V_1,V_2,V_3) \in \mathbb{R}^{m_1\times m_2\times m_3}$ is

$$[T(V_1,V_2,V_3)]_{i_1,i_2,i_3} := \sum_{j_1,j_2,j_3\in[d]} T_{j_1,j_2,j_3}\ [V_1]_{j_1,i_1}\ [V_2]_{j_2,i_2}\ [V_3]_{j_3,i_3}.$$

The term multilinear map arises from the fact that the above map is linear in each of the coordinates, e.g. if we replace $V_1$ by $aV_1 + bW_1$ in the above equation, where $W_1$ is a matrix of appropriate dimensions, and $a, b$ are any scalars, the output is a linear combination of the outputs under $V_1$ and $W_1$ respectively. We will use the above notion of multi-linear transforms to describe various tensor operations. For instance, $T(I,I,v)$ yields a matrix, $T(I,v,v)$, a vector, and $T(v,v,v)$, a scalar.

**Symmetric tensors and orthogonal decomposition:** A special class of tensors are the symmetric tensors $T \in \mathbb{R}^{d\times d\times d}$ which are invariant to permutation of the array indices. Symmetric tensors have CP decomposition of the form

$$T = \sum_{i\in[r]} \lambda_i v_i \otimes v_i \otimes v_i = \sum_{i\in[r]} \lambda_i v_i^{\otimes 3}, \tag{16}$$

where $r$ denotes the tensor CP rank and we use the notation $v_i^{\otimes 3} := v_i \otimes v_i \otimes v_i$. It is convenient to first analyze methods for decomposition of symmetric tensors and we then extend them to the general case of asymmetric tensors.

Further, a sub-class of symmetric tensors are those which possess a decomposition into orthogonal components, i.e. the vectors $v_i \in \mathbb{R}^d$ are orthogonal to one another in the above decomposition in (16) (without loss of generality, we assume that vectors $\{v_i\}$ are orthonormal in this case). An orthogonal decomposition implies that the tensor rank $r \le d$ and there are tractable methods for recovering the rank-one components in this setting. We limit ourselves to this setting in this paper.

**Tensor eigen analysis:** For symmetric tensors $T$ possessing an orthogonal decomposition of the form in (16), each pair $(\lambda_i, v_i)$, for $i \in [r]$, can be interpreted as an eigen-pair for the tensor $T$, since

$$T(I, v_i, v_i) = \sum_{j \in [r]} \lambda_j \langle v_i, v_j \rangle^2 v_j = \lambda_i v_i, \quad \forall i \in [r],$$

due to the fact that $\langle v_i, v_j \rangle = \delta_{i,j}$. Thus, the vectors $\{v_i\}_{i \in [r]}$ can be interpreted as fixed points of the map

$$v \mapsto \frac{T(I, v, v)}{\|T(I, v, v)\|}, \tag{17}$$

where $\|\cdot\|$ denotes the spectral norm (and $\|T(I, v, v)\|$ is a vector norm), and is used to normalize the vector $v$ in (17).

**Basic tensor power iteration method:** A straightforward approach to computing the orthogonal decomposition of a symmetric tensor is to iterate according to the fixed-point map in (17) with an arbitrary initialization vector. This is referred to as the tensor power iteration method. Additionally, it is known that the vectors $\{v_i\}_{i \in [r]}$ are the only *stable* fixed points of the map in (17). In other words, the set of initialization vectors which converge to vectors other than $\{v_i\}_{i \in [r]}$ are of measure zero. This ensures that we obtain the correct set of vectors through power iterations and that no spurious answers are obtained. See [4, Thm. 4.1] for details. Moreover, after an approximately fixed point is obtained (after many power iterations), the estimated eigen-pair can be subtracted out (i.e., *deflated*) and subsequent vectors can be similarly obtained through power iterations. Thus, we can obtain all the stable eigen-pairs $\{\lambda_i, v_i\}_{i \in [r]}$ which are the components of the orthogonal tensor decomposition. The method needs to be suitably modified when the tensor $T$ is perturbed (e.g. as in the case when empirical moments are used) and we discuss it in Section 3.3.

## 3.2 Learning Mixed Membership Models Under Exact Moments

We first describe the learning approach when exact moments are available. In Section 3.3, we suitably modify the approach to handle perturbations, which are introduced when only empirical moments are available.

We now employ the tensor power method described above to obtain a CP decomposition of the graph moment tensor $\mathrm{T}^{\alpha_0}$ in (12). We first describe a "symmetrization" procedure to convert the graph moment tensor $\mathrm{T}^{\alpha_0}$ to a symmetric orthogonal tensor through a multi-linear transformation of $\mathrm{T}^{\alpha_0}$. We then employ the power method to obtain a symmetric orthogonal decomposition. Finally, the original CP decomposition is obtained by reversing the multi-linear transform of the symmetrization procedure. This yields a guaranteed method for obtaining the decomposition of graph moment tensor $\mathrm{T}^{\alpha_0}$ under exact moments. We note that this symmetrization approach has been earlier employed in other contexts, e.g. for learning hidden Markov models [4, Sec. 3.3].

**Reduction of the graph-moment tensor to symmetric orthogonal form (Whitening):** Recall from Proposition 2.2 that the modified 3-star count tensor $\mathrm{T}^{\alpha_0}$ has a CP decomposition as

$$\mathbb{E}[\mathrm{T}^{\alpha_0} | \Pi_A, \Pi_B, \Pi_C] = \sum_{i=1}^{k} \widehat{\alpha}_i (F_A)_i \otimes (F_B)_i \otimes (F_C)_i.$$

We now describe a symmetrization procedure to convert $\mathrm{T}^{\alpha_0}$ to a symmetric orthogonal tensor through a multi-linear transformation using the modified adjacency matrix $G^{\alpha_0}$, defined in (11). Consider the singular value decomposition (SVD) of the modified adjacency matrix $G^{\alpha_0}$ under exact moments:

$$|X|^{-1/2}\mathbb{E}[(G^{\alpha_0}_{X,A})^\top|\Pi] = U_A D_A V_A^\top.$$

Define $W_A := U_A D_A^{-1}$, and similarly define $W_B$ and $W_C$ using the corresponding matrices $G^{\alpha_0}_{X,B}$ and $G^{\alpha_0}_{X,C}$ respectively. Now define

$$R_{A,B} := \frac{1}{|X|} W_B^\top \mathbb{E}[(G^{\alpha_0}_{X,B})^\top|\Pi] \cdot \mathbb{E}[(G^{\alpha_0}_{X,A})|\Pi] W_A, \quad \tilde{W}_B := W_B R_{A,B}, \tag{18}$$

and similarly define $\tilde{W}_C$. We establish that a multilinear transformation (as defined in (3.1)) of the graph-moment tensor $\mathrm{T}^{\alpha_0}$ using matrices $W_A, \tilde{W}_B$, and $\tilde{W}_C$ results in a symmetric orthogonal form.

**Lemma 3.1** (Orthogonal Symmetric Tensor)**.** *Assume that the matrices $F_A, F_B, F_C$ and $\Pi_X$ have rank $k$, where $k$ is the number of communities. We have an orthogonal symmetric tensor form for the modified* 3*-star count tensor $\mathrm{T}^{\alpha_0}$ in* (12) *under a multilinear transformation using matrices $W_A, \tilde{W}_B$, and $\tilde{W}_C$:*

$$\mathbb{E}[\mathrm{T}^{\alpha_0}(W_A, \tilde{W}_B, \tilde{W}_C)|\Pi_A, \Pi_B, \Pi_C] = \sum_{i\in[k]} \lambda_i (\Phi)_i^{\otimes 3} \in \mathbb{R}^{k\times k\times k}, \tag{19}$$

*where $\lambda_i := \widehat{\alpha}_i^{-0.5}$ and $\Phi \in \mathbb{R}^{k\times k}$ is an orthogonal matrix, given by*

$$\Phi := W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{0.5}). \tag{20}$$

**Remark 1:** Note that the matrix $W_A$ orthogonalizes $F_A$ under exact moments, and is referred to as a *whitening matrix*. Similarly, the matrices $\tilde{W}_B = R_{A,B} W_B$ and $\tilde{W}_C = R_{A,C} W_C$ consist of whitening matrices $W_B$ and $W_C$, and in addition, the matrices $R_{A,B}$ and $R_{A,C}$ serve to symmetrize the tensor. We can interpret $\{\lambda_i, (\Phi)_i\}_{i\in[k]}$ as the stable eigen-pairs of the transformed tensor (henceforth, referred to as the *whitened and symmetrized tensor*).

**Remark 2:** The full rank assumption on matrix $F_A = \Pi_A^\top P^\top \in \mathbb{R}^{|A|\times k}$ implies that $|A| \geq k$, and similarly $|B|, |C|, |X| \geq k$. Moreover, we require the community connectivity matrix $P \in \mathbb{R}^{k\times k}$ to be of full rank (which is a natural non-degeneracy condition). In this case, we can reduce the graph-moment tensor $\mathrm{T}^{\alpha_0}$ to a $k$-rank orthogonal symmetric tensor, which has a unique decomposition. This implies that the mixed membership model is identifiable using 3-star and edge count moments, when the network size $n = |A|+|B|+|C|+|X| \geq 4k$, matrix $P$ is full rank and the community membership matrices $\Pi_A, \Pi_B, \Pi_C, \Pi_X$ each have rank $k$. On the other hand, when only empirical moments are available, roughly, we require the network size $n = \Omega(k^2(\alpha_0+1)^2)$ (where $\alpha_0 := \sum_i \alpha_i$ is related to the extent of overlap between the communities) to provide guaranteed learning of the community membership and model parameters. See Section 4 for a detailed sample analysis.

*Proof:* Recall that the modified adjacency matrix $G^{\alpha_0}$ satisfies

$$\mathbb{E}[(G^{\alpha_0}_{X,A})^\top|\Pi_A, \Pi_X] = F_A \operatorname{Diag}(\widehat{\alpha}^{1/2})\Psi_X.$$
$$\Psi_X := \operatorname{Diag}(\widehat{\alpha}^{-1/2}) \left(\sqrt{\alpha_0+1}\Pi_X - (\sqrt{\alpha_0+1}-1)\left(\frac{1}{|X|}\sum_{i\in X}\pi_i\right)\vec{1}^\top\right).$$

From the definition of $\Psi_X$ above, we see that it has rank $k$ when $\Pi_X$ has rank $k$. Using the Sylvester's rank inequality, we have that the rank of $F_A \operatorname{Diag}(\widehat{\alpha}^{1/2})\Psi_X$ is at least $2k-k=k$. This implies that the whitening matrix $W_A$ also has rank $k$. Notice that

$$|X|^{-1} W_A^\top \mathbb{E}[(G^{\alpha_0}_{X,A})^\top|\Pi] \cdot \mathbb{E}[(G^{\alpha_0}_{X,A})|\Pi] W_A = D_A^{-1} U_A^\top U_A D_A^2 U_A^\top U_A D_A^{-1} = I \in \mathbb{R}^{k\times k},$$

or in other words, $|X|^{-1} M M^\top = I$, where $M := W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{1/2}) \Psi_X$. We now have that

$$I = |X|^{-1} \mathbb{E}_{\Pi_X}\left[ M M^\top \right] = |X|^{-1} W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{1/2}) \mathbb{E}[\Psi_X \Psi_X^\top] \operatorname{Diag}(\widehat{\alpha}^{1/2}) F_A^\top W_A$$
$$= W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}) F_A^\top W_A,$$

since $|X|^{-1}\mathbb{E}_{\Pi_X}[\Psi_X \Psi_X^\top] = I$ from (15), and we use the fact that the sets $A$ and $X$ do not overlap. Thus, $W_A$ whitens $F_A \operatorname{Diag}(\widehat{\alpha}^{1/2})$ under exact moments (up on taking expectation over $\Pi_X$) and the columns of $W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{1/2})$ are orthonormal. Now note from the definition of $\tilde{W}_B$ that

$$\tilde{W}_B^\top \mathbb{E}[(G_{X,B}^{\alpha_0})^\top | \Pi] = W_A^\top \mathbb{E}[(G_{X,A}^{\alpha_0})^\top | \Pi],$$

since $W_B$ satisfies

$$|X|^{-1} W_B^\top \mathbb{E}[(G_{X,B}^{\alpha_0})^\top | \Pi] \cdot \mathbb{E}[(G_{X,B}^{\alpha_0}) | \Pi] W_B = I,$$

and similar result holds for $\tilde{W}_C$. The final result in (19) follows by taking expectation of tensor $\mathrm{T}^{\alpha_0}$ over $\Pi_X$. □

**Overview of the learning approach under exact moments:** With the above result in place, we are now ready to describe the high-level approach for learning the mixed membership model under exact moments. First, symmetrize the graph-moment tensor $\mathrm{T}^{\alpha_0}$ as described above and then apply the tensor power method described in the previous section. This enables us to obtain the vector of eigenvalues $\lambda := \widehat{\alpha}^{-1/2}$ and the matrix of eigenvectors $\Phi = W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{0.5})$ using tensor power iterations. We can then recover the community membership vectors of set $A^c$ (i.e., nodes not in set $A$) under exact moments as

$$\Pi_{A^c} \leftarrow \operatorname{Diag}(\lambda)^{-1} \Phi^\top W_A^\top \mathbb{E}[G_{A^c,A}^\top | \Pi],$$

since $\mathbb{E}[G_{A^c,A}^\top | \Pi] = F_A \Pi_{A^c}$ (since $A$ and $A^c$ do not overlap) and $\operatorname{Diag}(\lambda)^{-1}\Phi^\top W_A^\top = \operatorname{Diag}(\widehat{\alpha}) F_A^\top W_A W_A^\top$ under exact moments. In order to recover the community membership vectors of set $A$, viz., $\Pi_A$, we can reverse the direction of the 3-star counts, i.e., consider the 3-stars from set $A$ to $X, B, C$ and obtain $\Pi_A$ in a similar manner. Once all the community membership vectors $\Pi$ are obtained, we can obtain the community connectivity matrix $P$, using the relationship: $\Pi^\top P \Pi = \mathbb{E}[G|\Pi]$ and noting that we assume $\Pi$ to be of rank $k$. Thus, we are able to learn the community membership vectors $\Pi$ and the model parameters $\widehat{\alpha}$ and $P$ of the mixed membership model using edge counts and the 3-star count tensor. We now describe modifications to this approach to handle empirical moments.

## 3.3 Learning Algorithm Under Empirical Moments

In the previous section, we explored a tensor-based approach for learning mixed membership model under exact moments. However, in practice, we only have samples (i.e. the observed network), and the method needs to be robust to perturbations when empirical moments are employed.

### 3.3.1 Pre-processing steps

**Partitioning:** In the previous section, we partitioned the nodes into four sets $A, B, C, X$ for learning under exact moments. However, we require more partitions under empirical moments to avoid statistical dependency issues and obtain stronger reconstruction guarantees. We now divide the network into five non-overlapping sets $A, B, C, X, Y$. The set $X$ is employed to compute whitening matrices $\hat{W}_A$, $\hat{W}_B$ and $\hat{W}_C$, described in detail subsequently, the set $Y$ is employed to compute the 3-star count tensor $\mathrm{T}^{\alpha_0}$ and sets $A, B, C$ contain the leaves of the 3-stars under consideration. The roles of the sets can be interchanged to obtain the community membership vectors of all the sets.

**Algorithm 1** $\{\hat{\Pi}, \hat{P}, \widehat{\alpha}\} \leftarrow \text{LearnMixedMembership}(G, k, \alpha_0, N, \tau)$

**Input:** Adjacency matrix $G \in \mathbb{R}^{n\times n}$, $k$ is the number of communities, $\alpha_0 := \sum_i \alpha_i$, where $\alpha$ is the Dirichlet parameter vector, $N$ is the number of iterations for the tensor power method, and $\tau$ is used for thresholding the estimated community membership vectors . Let $A^c := [n] \setminus A$ denote the set of nodes not in $A$.

**Output:** Estimates of the community membership vectors $\Pi \in \mathbb{R}^{n\times k}$, community connectivity matrix $P \in [0,1]^{k\times k}$, and the normalized Dirichlet parameter vector $\widehat{\alpha}$.

Partition the graph $G$ into 5 parts $X$, $Y$, $A$, $B$, $C$.

Compute moments $G^{\alpha_0}_{X,A}$, $G^{\alpha_0}_{X,B}$, $G^{\alpha_0}_{X,C}$, $\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}$ using (11) and (12).

$\{\hat{\Pi}_{A^c}, \widehat{\alpha}\} \leftarrow \text{LearnPartitionCommunity}(G^{\alpha_0}_{X,A},\ G^{\alpha_0}_{X,B},\ G^{\alpha_0}_{X,C},\ \mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}, G, N, \tau)$.

Interchange roles[10] of $Y$ and $A$ to obtain $\hat{\Pi}_{Y^c}$.

Define $\hat{Q} := \frac{\alpha_0+1}{n}\left(\hat{\Pi}\,\mathrm{Diag}(\widehat{\alpha}^{-1}) - \frac{\alpha_0}{\alpha_0+1}\vec{1}\vec{1}^\top\right)$

Estimate $\hat{P} \leftarrow \hat{Q}G\hat{Q}^\top$.

Return $\hat{\Pi}, \hat{P}, \widehat{\alpha}$

**Procedure 1** $\{\hat{\Pi}_{A^c}, \widehat{\alpha}\} \leftarrow \text{LearnPartitionCommunity}(G^{\alpha_0}_{X,A},\ G^{\alpha_0}_{X,B},\ G^{\alpha_0}_{X,C},\ \ \mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}},\ \ G,\ N,\ \tau)$

**Input:** Require modified adjacency submatrices $G^{\alpha_0}_{X,A}$, $G^{\alpha_0}_{X,B}$, $G^{\alpha_0}_{X,C}$, 3-star count tensor $\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}$, adjacency matrix $G$, number of iterations $N$ for the tensor power method and threshold $\tau$ for thresholding estimated community membership vectors. Let $\mathrm{Thres}(A,\tau)$ denote the element-wise thresholding operation using threshold $\tau$, i.e., $\mathrm{Thres}(A,\tau)_{i,j} = A_{i,j}$ if $A_{i,j} \geq \tau$ and 0 otherwise. Let $e_i$ denote basis vector along coordinate $i$.

**Output:** Estimates of $\Pi_{A^c}$ and $\widehat{\alpha}$.

Compute rank-$k$ SVD: $(G_{X,A})_{k-svd} = U_A D_A V_A^\top$ and compute whitening matrices $\hat{W}_A := U_A D_A^{-1}$. Similarly, compute $\hat{W}_B, \hat{W}_C$ and $\hat{R}_{AB}, \hat{R}_{AC}$ using (21).

Compute whitened and symmetrized tensor $T \leftarrow \mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC})$.

$\{\hat{\lambda}, \hat{\Phi}\} \leftarrow$TensorEigen$(T, \{\hat{W}_A^\top G_{i,A}^\top\}_{i\notin A}, N)$. $\{\hat{\Phi}$ is a $k\times k$ matrix with each columns being an estimated eigenvector and $\hat{\lambda}$ is the vector of estimated eigenvalues.$\}$

$\hat{\Pi}_{A^c} \leftarrow \mathrm{Thres}(\mathrm{Diag}(\hat{\lambda})^{-1}\hat{\Phi}^\top\hat{W}_A^\top G_{A^c,A}^\top\ ,\ \tau)$ and $\hat{\alpha}_i \leftarrow \hat{\lambda}_i^{-2}$, for $i \in [k]$.

Return $\hat{\Pi}_{A^c}$ and $\hat{\alpha}$.

**Whitening:** The whitening procedure is along the same lines as described in the previous section, except that now empirical moments are used. Specifically, consider the $k$-rank singular value decomposition (SVD) of the modified adjacency matrix $G^{\alpha_0}$ defined in (11),

$$|X|^{-1/2}(G^{\alpha_0}_{X,A})^\top_{k-svd} = U_A D_A V_A^\top.$$

Define $\hat{W}_A := U_A D_A^{-1}$, and similarly define $\hat{W}_B$ and $\hat{W}_C$ using the corresponding matrices $G^{\alpha_0}_{X,B}$ and $G^{\alpha_0}_{X,C}$ respectively. Now define

$$\hat{R}_{A,B} := \frac{1}{|X|}\hat{W}_B^\top (G^{\alpha_0}_{X,B})^\top_{k-svd} \cdot (G^{\alpha_0}_{X,A})_{k-svd}\hat{W}_A, \tag{21}$$

and similarly define $\hat{R}_{AC}$. The whitened and symmetrized graph-moment tensor is now computed as

$$\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC}),$$

where $\mathrm{T}^{\alpha_0}$ is given by (12) and the multi-linear transformation of a tensor is defined in (3.1).

#### 3.3.2 Modifications to the tensor power method

Recall that under exact moments, the stable eigen-pairs of a symmetric orthogonal tensor can be computed in a straightforward manner through the basic power iteration method in (17), along with the deflation procedure. However, this is not sufficient to get good reconstruction guarantees under empirical moments. We now propose a robust tensor method, detailed in Procedure 2. The main modifications involve: *(i)* efficient initialization and *(ii)* adaptive deflation, which are detailed below.

---

**Procedure 2** $\{\lambda, \Phi\} \leftarrow$TensorEigen($T$, $\{v_i\}_{i\in[L]}, N$)

---

**Input:** Tensor $T \in \mathbb{R}^{k\times k\times k}$, set of $L$ initialization vectors $\{v_i\}_{i\in L}$, number of iterations $N$.
**Output:** the estimated eigenvalue/eigenvector pairs $\{\lambda, \Phi\}$, where $\lambda$ is the vector of eigenvalues and $\Phi$ is the matrix of eigenvectors.

- **for** $i = 1$ to $k$ **do**
  - **for** $\tau = 1$ to $L$ **do**
    - $\theta_0 \leftarrow v_\tau$.
    - **for** $t = 1$ to $N$ **do**
      - $\tilde{T} \leftarrow T$.
      - **for** $j = 1$ to $i-1$ (when $i > 1$) **do**
        - **if** $|\lambda_j \left\langle \theta_t^{(\tau)}, \phi_j \right\rangle| > \xi$ **then**
          - $\tilde{T} \leftarrow \tilde{T} - \lambda_j \phi_j^{\otimes 3}$.
        - **end if**
      - **end for**
      - Compute power iteration update $\theta_t^{(\tau)} := \frac{\tilde{T}(I,\theta_{t-1}^{(\tau)},\theta_{t-1}^{(\tau)})}{\|\tilde{T}(I,\theta_{t-1}^{(\tau)},\theta_{t-1}^{(\tau)})\|}$
    - **end for**
  - **end for**
  - Let $\tau^* := \arg\max_{\tau\in L}\{\tilde{T}(\theta_N^{(\tau)}, \theta_N^{(\tau)}, \theta_N^{(\tau)})\}$.
  - Do $N$ power iteration updates starting from $\theta_N^{(\tau^*)}$ to obtain eigenvector estimate $\phi_i$, and set $\lambda_i := \tilde{T}(\phi_i, \phi_i, \phi_i)$.
- **end for**
- **return** the estimated eigenvalue/eigenvectors $(\lambda, \Phi)$.

---

**Efficient Initialization:** Recall that the basic tensor power method incorporates generic initialization vectors and this procedure recovers all the stable eigenvectors correctly (except for initialization vectors over a set of measure zero). However, under empirical moments, we have a perturbed tensor, and here, it is advantageous to instead employ specific initialization vectors. For instance, to obtain one of the eigenvectors $(\Phi)_i$, it is advantageous to initialize with a vector in the neighborhood of $(\Phi)_i$. This not only reduces the number of power iterations required to converge (approximately), but more importantly, this makes the power method more robust to perturbations. See Theorem B.1 in Appendix B.1 for a detailed analysis quantifying the relationship between initialization vectors, tensor perturbation and the resulting guarantees for recovery of the tensor eigenvectors.

For a mixed membership model in the sparse regime, recall that the community membership vectors $\Pi$ are sparse (with high probability). Under this regime of the model, we note that the whitened neighborhood vectors contain good initializers for the power iterations. Specifically, in Procedure 2, we initialize with the whitened neighborhood vectors $\hat{W}_A^\top G_{i,A}^\top$, for $i \notin A$. The intuition behind this is as follows: for a suitable choice of parameters (such as the scaling of network size $n$ with respect to the number of communities $k$), we expect neighborhood vectors $G_{i,A}^\top$ to concentrate around their mean values, viz., , $F_A \pi_i$. Since $\pi_i$ is sparse (w.h.p) for the model regime under consideration, this implies that there exist vectors $\hat{W}_A^\top F_A \pi_i$, for $i \in A^c$, which concentrate (w.h.p) on only along a few eigen-directions of the whitened tensor, and hence, serve as an effective initializer.

**Adaptive Deflation:** Recall that in the basic power iteration procedure, we can obtain the eigen-pairs one after another through simple deflation: subtracting the current estimates of eigen-pairs and running the power iterations again. However, it turns out that we can establish better robustness guarantees when we adaptively deflate the tensor in each power iteration. In Procedure 2, among the estimated eigen-pairs, we only deflate those which "compete" with the current estimate of the power iteration. In other words, if the vector in the current iteration has a significant projection along the direction of an estimated eigen-pair, then the eigen-pair is deflated; otherwise it is retained and not deflated. This allows us to carefully control the error build-up for each estimated eigenpair in our analysis. See Theorem B.1 in Appendix B.1 for details.

In addition, we note that stabilization, as proposed in [28] for general tensor eigen-decomposition (as opposed to orthogonal decomposition in this paper), can be effective in improving convergence, especially on real data, and we defer its detailed analysis to future work.

### 3.3.3 Reconstruction after tensor power method

Recall that previously in Section 3.2, when exact moments are available, estimating the community membership vectors $\Pi$ is straightforward, once we recover all the stable tensor eigen-pairs. However, in case of empirical moments, we can obtain better guarantees with the following modification: the estimated community membership vectors $\Pi$ are further subject to thresholding so that the weak values are set to zero. Since we are limiting ourselves to the regime of the mixed membership model, where the community vectors $\Pi$ are sparse (w.h.p), this modification strengthens our reconstruction guarantees. This thresholding step is incorporated in Algorithm 1.

Moreover, recall that under exact moments, estimating the community connectivity matrix $P$ is straightforward, once we recover the community membership vectors since $P \leftarrow (\Pi^\top)^\dagger \mathbb{E}[G|\Pi]\Pi^\dagger$. However, when empirical moments are available, we are able to establish better reconstruction guarantees through a different method, outlined in Algorithm 1. We define

$$\hat{Q} := \frac{\alpha_0 + 1}{n} \left( \hat{\Pi} \operatorname{Diag}(\widehat{\alpha}^{-1}) - \frac{\alpha_0}{\alpha_0 + 1} \vec{1}\vec{1}^\top \right),$$

based on estimates $\hat{\Pi}$ and $\widehat{\alpha}$, and the matrix $\hat{P}$ is obtained as $\hat{P} \leftarrow QGQ^\top$. Note that under exact moments, we have $\hat{\Pi} = \Pi$, i.e., community vectors are recovered perfectly, and

$$\mathbb{E}_\Pi[Q\Pi^\top] = \frac{\alpha_0 + 1}{n} \left( \mathbb{E}[\Pi \operatorname{Diag}(\lambda^2)\Pi^\top] - \frac{\alpha_0}{\alpha_0 + 1} \vec{1}\vec{1}^\top \mathbb{E}[\Pi^\top] \right) = I,$$

where $Q$ is the counterpart of $\hat{Q}$ under exact moments. Thus, we see that in the case of exact moments, the community connectivity matrix $P$ is recovered exactly. We now carry out careful sample analysis for the above method and the results are summarized in the next section.

**Improved support recovery estimates in homophilic models:** A sub-class of community model are those satisfying *homophily*. As discussed in Section 1, homophily or the tendency to form edges within the members of the same community has been posited as an important factor in community formation, especially in social settings. Many of the existing learning algorithms, e.g. [44], require this assumption to provide guarantees in the stochastic block model setting.

Specifically, we require models with community connectivity matrix $P \in [0,1]^{k\times k}$ satisfying $P(i,i) > P(i,j)$ for all $i \neq j$. For such models, we can obtain improved estimates by averaging, and is detailed in Procedure 3. Specifically, consider nodes in set $C$ and edges going from $C$ to nodes in $B$. First, consider the special case of the stochastic block model: for each node $c \in C$, compute the number of neighbors in $B$ belonging to each community (as given by the estimate $\hat{\Pi}$ from Algorithm 1), and declare the community with the maximum number of such neighbors as the community of node $c$. Intuitively, this provides a better estimate for $\Pi_C$ since we average over the edges in $B$. This method has been used before in the context of spectral clustering [32]. The same idea can be extended to the general mixed membership (homophilic) models: declare communities to be significant if they exceed a certain threshold, as evaluated by the average number of edges to each community. The details are provided in Procedure 3. In the next section, we establish that in certain regime of parameters, this procedure can lead to zero-error support recovery of significant community memberships of the nodes and also rule out communities where a node does not have a strong presence.

---

**Procedure 3** $\{\hat{S}\} \leftarrow$ SupportRecoveryHomophilicModels$(G, k, \alpha_0, \xi, \hat{\Pi})$

---

**Input:** Adjacency matrix $G \in \mathbb{R}^{n\times n}$, $k$ is the number of communities, $\alpha_0 := \sum_i \alpha_i$, where $\alpha$ is the Dirichlet parameter vector, $\xi$ is the threshold for support recovery, corresponding to significant community memberships of an individual. Get estimate $\hat{\Pi}$ from Algorithm 1. Also specify if the model is homophilic: whether $P(i,i) > P(i,j)$, for all $i \neq j$.

**Output:** $\hat{S} \in \{0,1\}^{n\times k}$ is the matrix supported on large entries of $\hat{\Pi}$, i.e. declare $\hat{S}(i,j) = 1$ if the (revised) estimate of $\Pi(i,j) \geq \xi$, and 0 otherwise.

**if** Model satisfies homophily **then**
  {Now provide improved estimates for support recovery in homophilic models}
  Consider partitions $A, B, C, X, Y$ as in Algorithm 1.
  Define $\hat{Q}^i_B := (\alpha_0+1)\frac{\hat{\Pi}^i_B}{|\hat{\Pi}^i_B|_1} + \frac{\alpha_0+1}{|B|}\vec{1}^\top$, similarly define $\hat{Q}^C$ {Define $\hat{Q}$ using $\hat{\Pi}$ from Algorithm 1.}
  Estimate $\hat{F}_C \leftarrow G_{C,B}\hat{Q}_B^\top$, $\hat{P} \leftarrow \hat{Q}_C\hat{F}_C$. {Define a new estimate for $P$.}
  **if** $\alpha_0 = 0$ (stochastic block model) **then**
    **for** $c \in C$ **do**
      Let $i^* \leftarrow \arg\max_{i\in[k]} \hat{F}_C(c,i)$ and $\hat{\Pi}_c \leftarrow e_{i^*}$. {Assign community with maximum average degree.}
    **end for**
  **else**
    Let $H$ be the average of diagonals of $\hat{P}$, $L$ be the average of off-diagonals of $\hat{P}$
    **for** $c \in C$, $i \in [k]$ **do**
      $\hat{S}(i,c) \leftarrow 1$ if $\hat{F}_C(c,i) \geq L + (H-L)\cdot\frac{3\xi}{4}$ and zero otherwise.{Identify large entries}
    **end for**
  **end if**
  Permute the roles of the matrices to get results for $A, B, X, Y$.
**else**
  $\hat{S}(i,j) \leftarrow \mathbb{I}[\hat{\Pi}(i,j) \geq \xi]$. {For general models, use the original $\hat{\Pi}$ estimate for support recovery.}
**end if**

---

**Computational complexity:** We note that the computational complexity of the method is $O(n^2k + k^{4.43}\widehat{\alpha}_{\min}^{-1})$ when $\alpha_0 > 1$ and $O(n^2k)$ when $\alpha_0 < 1$. This is because the time for computing whitening matrices is dominated by SVD of the top $k$ singular vectors of $n \times n$ matrix, which takes $O(n^2k)$ time. We then compute the whitened tensor $T$ which requires time $O(n^2k + k^3n) = O(n^2k)$, since for each $i \in Y$, we multiply $G_{i,A}, G_{i,B}, G_{i,C}$ with the corresponding whitening matrices, and this step takes $O(nk)$ time. We then average this $k \times k \times k$ tensor over different nodes $i \in Y$ to the result, which takes $O(k^3)$ time in each step.

For the tensor power method, the time required for a single iteration is $O(k^3)$. We need at most $\log n$ iterations per initial vector, and we need to consider $O(\widehat{\alpha}_{\min}^{-1}k^{0.43})$ initial vectors (this could be smaller when $\alpha_0 < 1$). Hence the total running time of tensor power method is $O(k^{4.43}\widehat{\alpha}_{\min}^{-1})$ (and when $\alpha_0$ is small this can be improved to $O(k^4\widehat{\alpha}_{\min}^{-1})$ which is dominated by $O(n^2k)$.

In the process of estimating $\Pi$ and $P$, the dominant operation is multiplying $k \times n$ matrix by $n \times n$ matrix, which takes $O(n^2k)$ time. For support recovery, the dominant operation is computing the "average degree", which again takes $O(n^2k)$ time. Thus, we have that the overall computational time is $O(n^2k + k^{4.43}\widehat{\alpha}_{\min}^{-1})$ when $\alpha_0 > 1$ and $O(n^2k)$ when $\alpha_0 < 1$.

# 4 Sample Analysis for Proposed Learning Algorithm

## 4.1 Sufficient Conditions and Recovery Guarantees

We now provide recovery guarantees for the proposed learning algorithm under empirical moments under a sufficient set of conditions, involving scaling of various parameters such as network size $n$, number of communities $k$, concentration parameter $\alpha_0$ of the Dirichlet distribution (which is a measure of overlap of the communities) and so on.

**[A1] Sparse regime of Dirichlet parameters:** The community membership vectors are drawn from the Dirichlet distribution, $\mathrm{Dir}(\alpha)$, under the mixed membership model. We assume that[11] $\alpha_i < 1$ for $i \in [k]$ $\alpha_i < 1$ for all $i \in [k]$ (see Section 2.1 for an extended discussion on the sparse regime of the Dirichlet distribution).

**[A2] Condition on the network size:** Given the concentration parameter of the Dirichlet distribution, $\alpha_0 := \sum_i \alpha_i$, and $\widehat{\alpha}_{\min} := \alpha_{\min}/\alpha_0$, the expected size of the smallest community, define

$$\rho := \frac{\alpha_0 + 1}{\widehat{\alpha}_{\min}}. \tag{22}$$

Given $\delta \in (0, 1)$, we require that the network size scale as

$$n = \Omega\left(\rho^2 \log^2 \frac{k}{\delta}\right), \tag{23}$$

and that the partitions $A, B, C, X, Y$ are $\Theta(n)$. Note that from assumption A1, $\alpha_i < 1$ which implies that $\alpha_0 < k$. Thus, in the worst-case, when $\alpha_0 = \Theta(k)$, we require[12] $n = \tilde{\Omega}(k^4)$, assuming equal sizes: $\widehat{\alpha}_i = 1/k$, and in the best case, when $\alpha_0 = \Theta(1)$, we require $n = \tilde{\Omega}(k^2)$. The latter case includes the stochastic block model ($\alpha_0 = 0$), and thus, our results match the state-of-art bounds for learning stochastic block models. See Section 4.2 for an extended discussion.

[11] The assumption A1 that the Dirichlet distribution be in the sparse regime is not strictly needed. Our results can be extended to general Dirichlet distributions, but with worse scaling requirements on $n$. The dependence of $n$ is still polynomial in $\alpha_0$, i.e. we require $n = \tilde{\Omega}((\alpha_0 + 1)^c \widehat{\alpha}_{\min}^{-2})$, where $c \geq 2$ is some constant.

[12] The notation $\tilde{\Omega}(\cdot), \tilde{O}(\cdot)$ denotes $\Omega(\cdot), O(\cdot)$ up to log factors.

**[A3] Condition on relative community sizes and block connectivity matrix:** Recall that $P \in [0,1]^{k\times k}$ denotes the block connectivity matrix. Define

$$\zeta := \left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2} \frac{\sqrt{(\max_i(P\widehat{\alpha})_i)}}{\sigma_{\min}(P)}, \tag{24}$$

where $\sigma_{\min}(P)$ is the minimum singular value of $P$. We require that

$$\zeta = \begin{cases} O\left(\dfrac{n^{1/2}}{\rho}\right), & \alpha_0 < 1 \qquad (25) \\ O\left(\dfrac{n^{1/2}}{\rho k \widehat{\alpha}_{\max}}\right) & \alpha_0 \geq 1. \qquad (26) \end{cases}$$

Intuitively, the above condition requires the ratio of maximum and minimum expected community sizes to be not too large and for the matrix $P$ to be well conditioned. The above condition is required to control the perturbation in the whitened tensor (computed using observed network samples), thereby, providing guarantees on the estimated eigen-pairs through the tensor power method. The above condition can be interpreted as a separation requirement between intra-community and inter-community connectivity in the special case considered in Section 4.2.

**[A4] Condition on number of iterations of the power method:** We assume that the number of iterations $N$ of the tensor power method in Procedure 2 satisfies

$$N \geq C_2 \cdot \left(\log(k) + \log\log\left(\frac{\sigma_{\min}(P)}{(\max_i(P\widehat{\alpha})_i)}\right)\right), \tag{27}$$

for some constant $C_2$.

**[A5] Choice of $\tau$ for thresholding community vector estimates:** The threshold $\tau$ for obtaining estimates $\hat{\Pi}$ of community membership vectors in Algorithm 1 is chosen as

$$\tau = \begin{cases} \Theta\left(\dfrac{\rho^{1/2}\cdot\zeta\cdot\widehat{\alpha}_{\max}^{1/2}}{n^{1/2}\cdot\widehat{\alpha}_{\min}}\right), & \alpha_0 \neq 0, \qquad (28) \\ 0.5, & \alpha_0 = 0, \qquad (29) \end{cases}$$

For the stochastic block model ($\alpha_0 = 0$), since $\pi_i$ is a basis vector, we can use a large threshold. For general models ($\alpha_0 \neq 0$), $\tau$ can be viewed as a regularization parameter and decays as $n^{-1/2}$ when other parameters are held fixed. Moreover, when $n = \tilde{\Theta}(\rho^2)$, we have that $\tau \sim \rho^{-1/2}$ when other terms are held fixed. Recall that $\rho \propto (\alpha_0 + 1)$ when the expected community sizes $\widehat{\alpha}_i$ are held fixed. In this case, $\tau \sim \rho^{-1/2}$ allows for smaller values to be picked up after thresholding as $\alpha_0$ is increased. This is intuitive since as $\alpha_0$ increases, the community vectors $\pi$ are more "spread out" across different communities and have smaller values.

We are now ready to state the error bounds on estimating $\Pi$ and $P$ using Algorithm 1. The proofs are given in the Appendix and a proof outline is provided in Section 4.3. Recall that for a matrix $M$, $(M)^i$ and $(M)_i$ denote the $i^{\text{th}}$ row and column respectively.

**Theorem 4.1** (Guarantees on estimating $P, \Pi$)**.** *Under assumptions A1-A5, The estimates $\hat{P}$ and $\hat{\Pi}$ obtained from Algorithm 1 satisfy with high probability,*

$$\varepsilon_{\pi,\ell_1} := \max_{i\in[k]} |(\hat{\Pi}_Z)^i - (\Pi_Z)^i|_1 = \tilde{O}\left(n^{1/2}\cdot\rho^{3/2}\cdot\zeta\cdot\widehat{\alpha}_{\max}\right) \tag{30}$$

$$\varepsilon_P := \max_{i,j\in[n]} |(\hat{Q}G\hat{Q}^\top)_{i,j} - P_{i,j}| = \tilde{O}\left(n^{-1/2}\cdot\rho^{5/2}\cdot\zeta\cdot\widehat{\alpha}_{\max}^{3/2}\right) \tag{31}$$

Thus the above result provides recovery guarantees on the estimates of $\Pi$ and $P$. The main ingredient in establishing the above result is the tensor concentration bound and additionally, recovery guarantees under the tensor power method in Procedure 2. We now provide these results below.

Recall that $F_A := \Pi_A^\top P^\top$ and $\Phi = W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{1/2})$ denotes the set of tensor eigenvectors under exact moments in (20), and $\hat{\Phi}$ is the set of estimated eigenvectors under empirical moments, obtained using Procedure 1. We establish the following guarantees.

**Lemma 4.2** (Perturbation bound for estimated eigen-pairs). *Under the assumptions A1-A4, the recovered eigenvector-eigenvalue pairs $(\hat{\Phi}_i, \hat{\lambda}_i)$ from the tensor power method in Procedure 2 satisfies with probability $1 - 109\delta$, for a permutation $\theta$, such that*

$$\max_{i\in[k]} \|\hat{\Phi}_i - \Phi_{\theta(i)}\| \leq 8\widehat{\alpha}_{\max}^{1/2}\varepsilon_T, \qquad \max_i |\lambda_i - \widehat{\alpha}_{\theta(i)}^{-1/2}| \leq 5\varepsilon_T, \tag{32}$$

*The tensor perturbation bound $\varepsilon_T$ is given by*

$$\begin{aligned} \varepsilon_T &:= \left\| \mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC}) - \mathbb{E}[\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(W_A, \tilde{W}_B, \tilde{W}_C)|\Pi_{A\cup B\cup C}] \right\| \\ &= \tilde{O}\left(\frac{\rho}{\sqrt{n}} \cdot \frac{\zeta}{\widehat{\alpha}_{\max}^{1/2}}\right), \end{aligned} \tag{33}$$

*where $\|T\|$ for a tensor $T$ refers to its spectral norm, $\rho$ is defined in* (22) *and $\zeta$ in* (24).

## 4.2 Special case: uniform community sizes and structured $P$

It is easier to interpret the results from the earlier section for the special case, where all the communities have the same expected size and the entries of the community connectivity matrix $P$ are equal on diagonal and off-diagonal locations:

$$\widehat{\alpha}_i \equiv \frac{1}{k}, \qquad P(i,j) = p\cdot\mathbb{I}(i=j) + q\cdot\mathbb{I}(i\neq j), \quad p \geq q. \tag{34}$$

In other words, the probability of an edge according to $P$ only depends on whether it is between two individuals of the same communities or between different communities. The above setting is well studied for stochastic block models ($\alpha_0 = 0$) and we compare our results with existing results for this setting.

In this setting, we have

$$\sigma_{\min}(P) = \Theta(p-q), \quad \max_i (P\widehat{\alpha})_i = \frac{p}{k} + (k-1)\frac{q}{k} \leq p.$$

Thus, the assumptions A2 and A3 in Section 4.1 reduce to

$$n = \tilde{\Omega}(k^2(\alpha_0+1)^2), \qquad \zeta = \Theta\left(\frac{\sqrt{p}}{p-q}\right) = O\left(\frac{n^{1/2}}{(\alpha_0+1)k}\right). \tag{35}$$

Thus, we obtain transparent conditions on scaling for $n$ and the separation $p - q$ between intra-community and inter-community connectivity. We now provide recovery guarantees for this setting.

**Corollary 4.3** (Uniform community sizes and structured $P$). *Under assumptions A1-A5 in Section 4.1 and* (34)*, we have with high probability*

$$\begin{aligned} \varepsilon_{\pi,\ell_1} &:= \max_i \|\hat{\Pi}^i - \Pi^i\|_1 = \tilde{O}\left(\frac{(\alpha_0+1)^{3/2}\sqrt{np}}{(p-q)}\right) \\ \varepsilon_P &:= \max_{i,j\in[n]} |(\hat{Q}G\hat{Q}^\top)_{i,j} - P_{i,j}| = \tilde{O}\left(\frac{(\alpha_0+1)^{3/2}k\sqrt{p}}{(p-q)\sqrt{n}}\right) \\ \varepsilon_T &= \tilde{O}\left(\frac{(\alpha_0+1)k^{3/2}\sqrt{p}}{(p-q)\sqrt{n}}\right), \end{aligned}$$

*where $\varepsilon_T$ is the (whitened) tensor perturbation bound defined in* (33).

Note that the assumption $p \geq q$ is not required for the above results in Corollary 4.3 to hold and we can replace $p$ by $\max(p, q)$ and $p - q$ with $|p - q|$ in the above bounds. However, note that we require the assumption that $p \geq q$ in Section 4.2.1 to provide improved guarantees for support recovery using Procedure 3.

**Stochastic block models ($\alpha_0 = 0$):** For stochastic block models, (35) reduces to

$$n = \tilde{\Omega}(k^2), \qquad \zeta = \Theta\left(\frac{\sqrt{p}}{p-q}\right) = O\left(\frac{n^{1/2}}{k}\right). \tag{36}$$

This matches with the best known scaling (up to poly-log factors), and was previously achieved via convex optimization in [44] for stochastic block models. However, our results in Corollary 4.3 do not provide zero error guarantees as in [44] since it assumes a general mixed membership model. We strengthen our results to provide zero-error guarantees in Section 4.2.1 and thus, match the scaling of [44] for stochastic block models. Moreover, we also provide zero-error support recovery guarantees for recovering significant memberships of nodes in mixed membership models in Section 4.2.1.

**Dependence on $\alpha_0$:** The guarantees degrade as $\alpha_0$ increases, which is intuitive since the extent of community overlap increases. The requirement for scaling of $n$ also grows as $\alpha_0$ increases. Note that the guarantees on $\varepsilon_\pi$ and $\varepsilon_P$ can be improved by assuming a more stringent scaling of $n$ with respect to $\alpha_0$, rather than the one specified by (35).

### 4.2.1 Zero-error guarantees for support recovery

Recall that we proposed Procedure 3 to provide improved support recovery estimates for homophilic models (where there are more expected edges within each community than to any community outside). We now provide guarantees for this method. We limit our analysis to the setting in (34) with uniform communities and structured matrix $P$. In principle, the analysis can be extended to more general homophilic models with suitable modifications to the method in Procedure 3.

We now specify the threshold $\xi$ for support recovery in Procedure 3.

**[A6] Choice of $\xi$ for support recovery:** We assume that the threshold $\xi$ in Procedure 3 satisfies

$$\xi = \Omega(\varepsilon_P),$$

where $\varepsilon_P$ is specified in Corollary 4.3. We now state the guarantees for support recovery.

**Theorem 4.4** (Support recovery guarantees)**.** *Assuming A1-A6 and* (34) *hold, the support recovery method in Procedure 3 has the following guarantees on the estimated support set $\hat{S}$: with high probability,*

$$\Pi(i,j) \geq \xi \Rightarrow \hat{S}(i,j) = 1 \quad \text{and} \quad \Pi(i,j) \leq \frac{\xi}{2} \Rightarrow \hat{S}(i,j) = 0, \quad \forall i \in [k], j \in [n], \tag{37}$$

*where $\Pi$ is the true community membership matrix.*

Thus, the above result guarantees that the Procedure 3 correctly recovers all the “large” entries of $\Pi$ and also correctly rules out all the “small” entries in $\Pi$. In other words, we can correctly infer all the significant memberships of each node and also rule out the set of communities where a node does not have a strong presence.

The only shortcoming of the above result is that there is a gap between the “large” and “small” values, and for an intermediate set of values (in $[\xi/2, \xi]$), we cannot guarantee correct inferences about the community memberships. Note this gap depends on $\varepsilon_P$, the error in estimating the $P$ matrix. This is intuitive, since as the error $\varepsilon_P$ decreases, we can infer the community memberships over a large range of values.

For the special case of stochastic block models (i.e. $\lim \alpha_0 \to 0$), we can improve the above result and give a zero error guarantee at all nodes (w.h.p). Note that we no longer require a threshold $\xi$ in this case, and only infer one community for each node.

**Corollary 4.5** (Zero error guarantee for block models)**.** *Assuming A1-A5 and* (34) *hold, the support recovery method in Procedure 3 correctly identifies the community memberships for all nodes with high probability in case of stochastic block models* ($\alpha_0 \to 0$)*.*

Thus, with the above result, we match the state-of-art results of [44] for stochastic block models in terms of scaling requirements and recovery guarantees.

## 4.3 Proof Outline

We now summarize the main techniques involved in proving Theorem 4.1. The details are in the Appendix. The main ingredient is the concentration of the adjacency matrix: since the edges are drawn independently conditioned on the community memberships, we establish that the adjacency matrix concentrates around its mean under the stated assumptions. See Appendix C.4 for details. With this in hand, we can then establish concentration of various quantities used by our learning algorithm.

**Step 1: Whitening matrices.** We first establish concentration bounds on the whitening matrices $\hat{W}_A$, $\hat{W}_B$, $\hat{W}_C$ computed using empirical moments, described in Section 3.3.1. With this in hand, we can approximately recover the span of matrix $F_A$ since $\hat{W}_A^\top F \operatorname{Diag}(\widehat{\alpha}_i)^{1/2}$ is a rotation matrix. The main technique employed is the Matrix Bernstein's inequality [40, thm. 1.4]. See Appendix C.2 for details.

**Step 2: Tensor concentration bounds** Recall that we use the whitening matrices to obtain a symmetric orthogonal tensor. We establish that the whitened and symmetrized tensor concentrates around its mean. This is done in several stages and we carefully control the tensor perturbation bounds. See Appendix C.1 for details.

**Step 3: Tensor power method analysis.** We analyze the performance of Procedure 2 under empirical moments. We employ the various improvements, detailed in Section 3.3.2 to establish guarantees on the recovered eigen-pairs. This includes coming up with a condition on the tensor perturbation bound, for the tensor power method to succeed. It also involves establishing that there exist good initializers for the power method among (whitened) neighborhood vectors. This allows us to obtain stronger guarantees for the tensor power method, compared to earlier analysis in [4]. This analysis is crucial for us to obtain state-of-art scaling bounds for guaranteed recovery (for the special case of stochastic block model). See Appendix B for details.

**Step 4: Thresholding of estimated community vectors** In Step 3, we provide guarantees for recovery of each eigenvector in $\ell_2$ norm. Direct application of this result only allows us to obtain $\ell_2$ norm bounds for row-wise recovery of the community matrix $\Pi$. In order to strengthen the result to an $\ell_1$ norm bound, we threshold the estimated $\Pi$ vectors. Here, we exploit the sparsity in Dirichlet draws and carefully control the contribution of weak entries in the vector. Finally, we establish perturbation bounds on $P$ through rather straightforward concentration bound arguments. See Appendix A.2 for details.

**Step 5: Support recovery guarantees.** It is convenient to consider the case of in stochastic block model here in the canonical setting of Section 4.2. Recall that Procedure 3 readjusts the community membership estimates based on degree averaging. For each vertex, if we count the average degree towards these "approximate communities", for the correct community the result is concentrated around value $p$ and for the wrong community the result is around value $q$. Therefore, we can correctly identify the community memberships of all the nodes, when $p - q$ is sufficiently large, as specified by (35). The argument can be easily extended to general mixed membership models. See Appendix A.3 for details.

## Acknowledgement

We thank Jure Leskovec for helpful discussions regarding various community models. Part of this work was done when AA and RG were visiting MSR New England. AA is supported in part by the NSF Award CCF-1219234, AFOSR Award FA9550-10-1-0310 and the ARO Award W911NF-12-1-0404.

## References

[1] Ittai Abraham, Shiri Chechik, David Kempe, and Aleksandrs Slivkins. Low-distortion inference of latent similarities from a multiplex social network. *CoRR*, abs/1202.0922, 2012.

[2] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, June 2008.

[3] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y. Liu. Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation, 2012. arXiv:1204.6703.

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for latent variable models, 2012.

[5] A. Anandkumar, D. Hsu, and S.M. Kakade. A Method of Moments for Mixture Models and Hidden Markov Models. In *Proc. of Conf. on Learning Theory*, June 2012.

[6] Sanjeev Arora, Rong Ge, Sushant Sachdeva, and Grant Schoenebeck. Finding overlapping communities in social networks: toward a rigorous approach. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, 2012.

[7] Maria-Florina Balcan, Christian Borgs, Mark Braverman, Jennifer T. Chayes, and Shang-Hua Teng. I like her more than you: Self-determined communities. *CoRR*, abs/1201.4899, 2012.

[8] P.J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[9] P.J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):38–59, 2011.

[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

[11] B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

[12] S. Charles Brubaker and Santosh S. Vempala. Random tensors and planted cliques. In *RANDOM*, 2009.

[13] S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. *Arxiv preprint arxiv:1102.2650*, 2011.

[14] S. Currarini, M.O. Jackson, and P. Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009.

[15] V. Feldman, E. Grigorescu, L. Reyzin, and S. Vempala. The complexity of statistical algorithms. *arXiv preprint arXiv:1201.1214*, 2012.

[16] S.E. Fienberg, M.M. Meyer, and S.S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the american Statistical association*, 80(389):51–67, 1985.

[17] O. Frank and D. Strauss. Markov graphs. *Journal of the american Statistical association*, 81(395):832–842, 1986.

[18] Alan M. Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[19] Alan M. Frieze and Ravi Kannan. A new approach to the planted clique problem. In *FSTTCS*, 2008.

[20] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[21] P. Gopalan, D. Mimno, S. Gerrish, M. Freedman, and D. Blei. Scalable inference of overlapping communities. In *Advances in Neural Information Processing Systems 25*, pages 2258–2266, 2012.

[22] C. Hillar and L.-H. Lim. Most tensor problems are NP hard, 2012.

[23] P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. *Social networks*, 5(2):109–137, 1983.

[24] P.W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76(373):33–50, 1981.

[25] A. Jalali, Y. Chen, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. *arXiv preprint arXiv:1104.4803*, 2011.

[26] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press., Cambridge, MA, 1994.

[27] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455, 2009.

[28] T. G. Kolda and J. R. Mayo. Shifted power method for computing tensor eigenpairs. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1095–1124, October 2011.

[29] P.F. Lazarsfeld, R.K. Merton, et al. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18(1):18–66, 1954.

[30] L. Lovász. Very large graphs. *Current Developments in Mathematics*, 2008:67–128, 2009.

[31] M. McPherson, L. Smith-Lovin, and J.M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[32] F. McSherry. Spectral partitioning of random graphs. In *FOCS*, 2001.

[33] J.L. Moreno. Who shall survive?: A new approach to the problem of human interrelations. 1934.

[34] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[35] K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society, London, A.*, page 71, 1894.

[36] A. Rinaldo, S.E. Fienberg, and Y. Zhou. On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics*, 3:446–484, 2009.

[37] T.A.B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

[38] G.W. Stewart and J. Sun. *Matrix perturbation theory*, volume 175. Academic press New York, 1990.

[39] M. Telgarsky. Dirichlet draws are sparse with high probability. *ArXiv:1301.4917*, 2012.

[40] J.A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[41] Y.J. Wang and G.Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

[42] H.C. White, S.A. Boorman, and R.L. Breiger. Social structure from multiple networks. i. blockmodels of roles and positions. *American journal of sociology*, pages 730–780, 1976.

[43] E.P. Xing, W. Fu, and L. Song. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 4(2):535–566, 2010.

[44] Chen Yudong, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. In *Advances in Neural Information Processing Systems 25*, 2012.

# A Proof of Theorem 4.1

## A.1 Proof of Lemma 4.2

We have the tensor perturbation bound in Theorem C.1 as follows: given $\delta \in (0,1)$ and $\rho := \frac{\alpha_0+1}{\widehat{\alpha}_{\min}}$ and

$$n = \Omega\left(\rho^2 \log^2 \frac{k}{\delta}\right), \tag{38}$$

and

$$\zeta := \left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2} \frac{\sqrt{(\max_i (P\widehat{\alpha})_i)}}{\sigma_{\min}(P)},$$

the following tensor perturbation bound holds with probability $1-100\delta$,

$$\begin{aligned}\varepsilon_T &:= \left\| \mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC}) - \mathbb{E}[\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(W_A, \tilde{W}_B, \tilde{W}_C)|\Pi_{A\cup B\cup C}] \right\| \\ &= O\left(\frac{(\alpha_0+1)\sqrt{(\max_i(P\widehat{\alpha})_i)}}{n^{1/2}\widehat{\alpha}^{3/2}_{\min}\sigma_{\min}(P)} \cdot \left(1 + \left(\frac{\rho^2}{n}\log^2\frac{k}{\delta}\right)^{1/4}\right)\sqrt{\frac{\log k}{\delta}}\right) \\ &= \tilde{O}\left(\frac{\rho}{\sqrt{n}} \cdot \frac{\zeta}{\widehat{\alpha}^{1/2}_{\max}}\right).\end{aligned} \tag{39}$$

From Theorem B.1, we require that the perturbation of the tensor be small enough according to

$$\varepsilon_T \le C_1 \widehat{\alpha}^{-1/2}_{\max} r_0^2, \tag{40}$$

for some constant $C_1$, in order to guarantee recovery of the eigen-pairs under the tensor power iteration method, when initialized with a $(\gamma, r_0)$ good vector.

By Lemma C.10, when $\zeta = O(\sqrt{n}r_0^2/\rho)$, we have good initial vectors. The requirement that $\varepsilon_T \le C_1\widehat{\alpha}^{-1/2}_{max} r_0^2$ turns out to be equivalent to $\zeta = O(\sqrt{n}r_0^2/\rho)$.

Therefore when $\zeta = O(\sqrt{n}r_0^2/\rho)$, the assumptions of Theorem B.1 are satisfied. Recall that $r_0^2 = O(1/\widehat{\alpha}_{max}k)$ when $\alpha_0 > 1$ and $r_0^2 = O(1)$ for $\alpha_0 \le 1$.

Additionally from Lemma C.10, in order to obtain good initialization vectors with probability $1-9\delta$ under Dirichlet distribution, we require that

$$n = \tilde{\Omega}\left(\alpha_{\min}^{-1} k^{0.43} \log(k/\delta)\right), \tag{41}$$

when $\alpha_0 > 1$, which is always satisfied since we are assuming $\widehat{\alpha}^{-2}_{min} < n$.

From Theorem B.1, we see that the tensor power method returns eigenvalue-vector pair $(\hat{\lambda}_i, \hat{\Phi}_i)$ such that there exists a permutation $\theta$ such that

$$\max_{i\in[k]} \|\hat{\Phi}_i - \Phi_{\theta(i)}\| \le 8\widehat{\alpha}^{1/2}_{\max}\varepsilon_T, \tag{42}$$

and

$$\max_i |\lambda_i - \widehat{\alpha}^{-1/2}_{\theta(i)}| \le 5\varepsilon_T. \tag{43}$$

## A.2 Reconstruction after tensor power method

Let $(M)^i$ and $(M)_i$ denote the $i^{\text{th}}$ row and $i^{\text{th}}$ column in matrix $M$ respectively. Let $Z \subset A^c$ denote any subset of nodes not in $A$, considered in Procedure LearnPartition Community. Define

$$\tilde{\Pi}_Z := \mathrm{Diag}(\lambda)^{-1}\Phi^\top \hat{W}_A^\top G_{Z,A}^\top.$$

Recall that the final estimate $\hat{\Pi}_Z$ is obtained by thresholding $\tilde{\Pi}_Z$ element-wise with threshold $\tau$ in Procedure 1. We first analyze perturbation of $\tilde{\Pi}_Z$.

**Lemma A.1** (Reconstruction Guarantees for $\tilde{\Pi}_Z$)**.** *Assuming Lemma 4.2 holds and the tensor power method recovers eigenvectors and eigenvalues up to the guaranteed errors, we have with probability* $1-122\delta$*,*

$$\begin{aligned}\varepsilon_\pi := \max_{i\in Z}\|(\tilde{\Pi}_Z)^i - (\Pi_Z)^i\| &= O\left(\varepsilon_T \widehat{\alpha}_{\max}^{1/2}\left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2}\|\Pi_Z\|\right),\\ &= O\left(\rho\cdot\zeta\cdot\widehat{\alpha}_{\max}^{1/2}\left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2}\right)\end{aligned}$$

*where* $\varepsilon_T$ *is given by* (39)*.*

*Proof:* We have $(\tilde{\Pi}_Z)^i = \lambda_i^{-1}((\Phi)_i)^\top \hat{W}_A^\top G_{Z,A}^\top$. We will now use perturbation bounds for each of the terms to get the result.

The first term is

$$\begin{aligned}&\|\operatorname{Diag}(\lambda_i)^{-1} - \operatorname{Diag}(\widehat{\alpha}_i^{1/2})\|\cdot\|\operatorname{Diag}(\widehat{\alpha}^{1/2})\tilde{F}_A^\top\|\cdot\|\tilde{F}_A\|\cdot\|\Pi_Z\|\\ &\le 5\varepsilon_T\widehat{\alpha}_{\max}\widehat{\alpha}_{\min}^{-1/2}(1+\varepsilon_1)^2\|\Pi_Z\|\end{aligned}$$

from the fact that $\|\operatorname{Diag}(\widehat{\alpha}^{1/2})\tilde{F}_A^\top\| \le 1+\varepsilon_1$, where $\varepsilon_1$ is given by (68). The second term is

$$\begin{aligned}&\|\operatorname{Diag}(\widehat{\alpha}^{1/2})\|\cdot\|(\Phi)_i - \widehat{\alpha}_i^{1/2}(\tilde{F}_A)_i\|\cdot\|\tilde{F}_A\|\cdot\|\Pi_Z\|\\ &\le 8\widehat{\alpha}_{\max}\varepsilon_T\widehat{\alpha}_{\min}^{-1/2}(1+\varepsilon_1)\|\Pi_Z\|\end{aligned}$$

The third term is

$$\begin{aligned}&\|\widehat{\alpha}_i^{1/2}\|\cdot\|(\hat{W}_A^\top - W_A^\top)F_A\Pi_Z\|\\ &\le \widehat{\alpha}_{\max}^{1/2}\widehat{\alpha}_{\min}^{-1/2}\|\Pi_Z\|\epsilon_W &(44)\\ &\le O\left(\left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2}\varepsilon_T\widehat{\alpha}_{\min}^{1/2}\|\Pi_Z\|\right), &(45)\end{aligned}$$

from Lemma C.2 and finally, we have

$$\begin{aligned}&\|\widehat{\alpha}_i^{1/2}\|\cdot\|W_A\|\cdot\|G_{Z,A}^\top - F_A\Pi_Z\|\\ &\le O\left(\widehat{\alpha}_{\max}^{1/2}\frac{\sqrt{\alpha_0+1}}{\widehat{\alpha}_{\min}\sigma_{\min}(P)}\sqrt{(\max_i(P\widehat{\alpha})_i)(1+\varepsilon_2+\varepsilon_3)\log\frac{k}{\delta}}\right) &(46)\\ &\le O\left(\left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2}\varepsilon_T\sqrt{\alpha_0+1}(1+\varepsilon_2+\varepsilon_3)\sqrt{\frac{\log k}{\delta}}\right) &(47)\end{aligned}$$

from Lemma C.7 and Lemma C.8.

The third term in (45) dominates the last term in (47) since $(\alpha_0+1)\log k/\delta < n\widehat{\alpha}_{\min}$ (due to assumption A2 on scaling of $n$). □

We now show that if we threshold the entries of $\tilde{\Pi}_Z$, the the resulting matrix $\hat{\Pi}_Z$ has rows close to those in $\Pi_Z$ in $\ell_1$ norm.

**Lemma A.2** (Guarantees after thresholding)**.** *For* $\hat{\Pi}_Z := \operatorname{Thres}(\tilde{\Pi}_Z, \tau)$*, where* $\tau$ *is the threshold, we have with probability* $1-2\delta$*, that*

$$\begin{aligned}\varepsilon_{\pi,\ell_1} := \max_{i\in[k]}|(\hat{\Pi}_Z)^i - (\Pi_Z)^i|_1 = O\Bigg(&\sqrt{n\eta}\,\varepsilon_\pi\sqrt{\log\frac{1}{2\tau}}\left(1-\sqrt{\frac{2\log(k/\delta)}{n\eta\log(1/2\tau)}}\right)\\ &+n\eta\tau + \sqrt{(n\eta+4\tau^2)\log\frac{k}{\delta}} + \frac{\varepsilon_\pi^2}{\tau}\Bigg),\end{aligned}$$

*where* $\eta = \widehat{\alpha}_{\max}$ *when* $\alpha_0 < 1$ *and* $\eta = \alpha_{\max}$ *when* $\alpha_0 \in [1, k)$*.*

**Remark 1:** The above guarantee on $\hat{\Pi}_Z$ is stronger than for $\tilde{\Pi}_Z$ in Lemma A.1 since this is an $\ell_1$ guarantee on the rows compared to $\ell_2$ guarantee on rows for $\tilde{\Pi}_Z$.

**Remark 2:** When $\tau$ is chosen as

$$\tau = \Theta(\frac{\varepsilon_\pi}{\sqrt{n\eta}}) = \Theta\left(\frac{\rho^{1/2}\cdot\zeta\cdot\widehat{\alpha}_{\max}^{1/2}}{n^{1/2}\cdot\widehat{\alpha}_{\min}}\right),$$

we have that

$$\begin{aligned}\max_{i\in[k]}|(\hat{\Pi}_Z)^i-(\Pi_Z)^i|_1 &= \tilde{O}\left(\sqrt{n\eta}\cdot\varepsilon_\pi\right)\\ &= \tilde{O}\left(n^{1/2}\cdot\rho^{3/2}\cdot\zeta\cdot\widehat{\alpha}_{\max}\right)\end{aligned}$$

*Proof:* Let $S_i := \{j : \hat{\Pi}_Z(i,j) > 2\tau\}$. For a vector $v$, let $v_S$ denote the sub-vector by considering entries in set $S$. We now have

$$|(\hat{\Pi}_Z)^i-(\Pi_Z)^i|_1 \le |(\hat{\Pi}_Z)^i_{S_i}-(\Pi_Z)^i_{S_i}|_1 + |(\Pi_Z)^i_{S_i^c}|_1 + |(\hat{\Pi}_Z)^i_{S_i^c}|_1$$

**Case $\alpha_0 < 1$:** From Lemma C.11, we have $\mathbb{P}[\Pi(i,j)\ge 2\tau] \le 8\widehat{\alpha}_i\log(1/2\tau)$. Since $\Pi(i,j)$ are independent for $j\in Z$, we have from multiplicative Chernoff bound [26, Thm 9.2], that with probability $1-\delta$,

$$\max_{i\in[k]}|S_i| < 8n\widehat{\alpha}_{\max}\log\left(\frac{1}{2\tau}\right)\left(1-\sqrt{\frac{2\log(k/\delta)}{n\widehat{\alpha}_i\log(1/2\tau)}}\right).$$

We have

$$|(\tilde{\Pi}_Z)^i_{S_i}-(\Pi_Z)^i_{S_i}|_1 \le \varepsilon_\pi|S_i|^{1/2},$$

and the $i^{\text{th}}$ rows of $\tilde{\Pi}_Z$ and $\hat{\Pi}_Z$ can differ on $S_i$, we have $|\tilde{\Pi}_Z(i,j)-\hat{\Pi}_Z(i,j)|\le\tau$, for $j\in S_i$, and number of such terms is at most $\varepsilon_\pi^2/\tau^2$. Thus,

$$|(\tilde{\Pi}_Z)^i_{S_i}-(\hat{\Pi}_Z)^i_{S_i}|_1 \le \frac{\varepsilon_\pi^2}{\tau}.$$

For the other term, from Lemma C.11, we have

$$\mathbb{E}[\Pi_Z(i,j)\cdot\delta(\Pi_Z(i,j)\le 2\tau)] \le \widehat{\alpha}_i(2\tau).$$

Applying Bernstein's bound we have with probability $1-\delta$

$$\max_{i\in[k]}\sum_{j\in Z}\Pi_Z(i,j)\cdot\delta(\Pi_Z(i,j)\le 2\tau) \le n\widehat{\alpha}_{\max}(2\tau) + \sqrt{2(n\widehat{\alpha}_{\max}+4\tau^2)\log\frac{k}{\delta}}.$$

For $\hat{\Pi}^i_{S_i^c}$, we further divide $S_i^c$ into $T_i$ and $U_i$, where $T_i := \{j : \tau/2 < \Pi_Z(i,j)\le 2\tau\}$ and $U_i := \{j : \Pi_Z(i,j)\le \tau/2\}$.

In the set $T_i$, using similar argument we know $|(\Pi_Z)^i_{T_i}-(\tilde{\Pi}_Z)^i_{T_i}|_1 \le O(\varepsilon_\pi\sqrt{n\widehat{\alpha}_{\max}\log 1/\tau})$, therefore

$$|\hat{\Pi}^i_{T_i}|_1 \le |\tilde{\Pi}^i_{T_i}|_1 \le |\Pi^i_{T_i}-\tilde{\Pi}^i_{T_i}|_1 + |\Pi^i_{S_i^c}|_1 \le O(\varepsilon_\pi\sqrt{n\widehat{\alpha}_{\max}\log 1/\tau}).$$

Finally, for index $j\in U_i$, in order for $\hat{\Pi}_Z(i,j)$ be positive, it is required that $\tilde{\Pi}_Z(i,j)-\Pi_Z(i,j)\ge\tau/2$. In this case, we have

$$|(\hat{\Pi}_Z)^i_{U_i}|_1 \le \frac{4}{\tau}\left\|(\tilde{\Pi}_Z)^i_{U_i}-\Pi^i_{U_i}\right\|^2 \le \frac{4\varepsilon_\pi^2}{\tau}.$$

**Case** $\alpha_0 \in [1, k)$**:** From Lemma C.11, we see that the results hold when we replace $\widehat{\alpha}_{\max}$ with $\alpha_{\max}$. $\square$

Finally we would like to use the community vectors $\Pi$ and the adjacency matrix $G$ to estimate the $P$ matrix.

**Lemma A.3** (Reconstruction of $P$)**.** *Let* $\hat{Q} = \frac{\alpha_0+1}{n}\left(\hat{\Pi}\,\mathrm{Diag}(\lambda)^2 - \frac{\alpha_0}{\alpha_0+1}\vec{1}\vec{1}^\top\right)$ *and $G$ is the adjacency matrix. When with probability* $1-5\delta$,

$$\varepsilon_P := \max_{i,j\in[n]} |(\hat{Q}G\hat{Q}^\top)_{i,j} - P_{i,j}| \le O\left(\frac{(\alpha_0+1)^{3/2}\varepsilon_\pi}{\sqrt{n}}\widehat{\alpha}_{\min}^{-1}\widehat{\alpha}_{\max}^{1/2}\log\frac{nk}{\delta}\right)$$

*Proof:* Let $Q = \frac{\alpha_0+1}{n}\left(\Pi\,\mathrm{Diag}(\widehat{\alpha})^{-1} - \frac{\alpha_0}{\alpha_0+1}\vec{1}\vec{1}^\top\right)$. The proof goes in three steps:

$$P \approx Q\Pi^\top P\Pi Q^\top \approx QGQ^\top \approx \hat{Q}G\hat{Q}^\top.$$

First we show with high probability $Q\Pi^\top$ is a $k\times k$ matrix that is very close to identity (note that $Q$ is constructed so that $\mathbb{E}_\Pi[Q\Pi^\top] = I$). For all $i \ne j$, $(Q\Pi^\top)_{i,j}$ has absolute value at most $O(\sqrt{\log(nk/\delta)\cdot\widehat{\alpha}_i/\widehat{\alpha}_j}/\sqrt{n})$ with probability $1-\delta$. This is because each entry in $Q\Pi^\top$ can be viewed as the sum of $n$ independent variables $(\frac{\alpha_0+1}{n}\frac{\Pi_{i,t}-\mathbb{E}[\Pi_{i,t}]}{\widehat{\alpha}_i}\Pi_{j,t})$ with mean 0, and variance bounded by $O(\sqrt{\widehat{\alpha}_i/\widehat{\alpha}_j}/n)$ (recall $\mathbb{E}[\Pi_{i,t}^2] = \widehat{\alpha}_i\cdot\frac{\alpha_i+1}{\alpha_0+1} \le 2\widehat{\alpha}_i/(\widehat{\alpha}_0+1)$). The bound follows from Bernstein's inequality and union bound. For diagonal entries, $(Q\Pi^\top)_{i,i}$ can also be viewed as the sum of $n$ independent variables $(\frac{\alpha_0+1}{n}\frac{\Pi_{i,t}-\mathbb{E}[\Pi_{i,t}]}{\widehat{\alpha}_i}\Pi_{i,t})$ with mean $1/n$ and variance $\widehat{\alpha}_i^{-1/2}/n$, therefore its difference with 1 is bounded by $\widehat{\alpha}_i^{-1/2}/\sqrt{n}$ with high probability. Using these bounds, the $i$-th row of $Q\Pi^\top$ is $e_i^\top + \Delta_i^\top$ where $\Delta_i$ is an error vector with $\ell_1$ norm bounded by $k\sqrt{\log(nk/\delta)\widehat{\alpha}_{max}/\widehat{\alpha}_{min}}/\sqrt{n}$ with probability $1-\delta$.

Now consider the matrix $Q\Pi^\top P\Pi Q^\top$. When entries of $Q\Pi^\top$ satisfy the bounds we derived above, we know the $i,j$-th entry of $Q\Pi^\top P\Pi Q^\top$ is close to $P_{i,j}$, with probability $1-\delta$,

$$\begin{aligned}(Q\Pi^\top P\Pi Q^\top)_{i,j} - P_{i,j} &= (e_i+\Delta_i)^\top P(e_j+\delta_j) - P_{i,j}\\ &= \Delta_i^\top P(e_j+\Delta_j) + e_i^\top P\Delta_j\\ &\le 2|\Delta_i|_1 + |\Delta_j|_1\\ &\le O\left(\frac{k\cdot\sqrt{\widehat{\alpha}_{max}/\widehat{\alpha}_{min}}}{\sqrt{n}}\sqrt{\log\frac{nk}{\delta}}\right),\end{aligned}$$

where we used the fact that $P_{i,j}\in[0,1]$.

Next we show $Q\Pi^\top P\Pi Q^\top$ is close to $QGQ^\top$. This holds due to the properties of Dirichlet distribution. Note that the $\ell_2$ norm of each row of $Q$ is bounded by $O(\sqrt{\log(nk/\delta)/(n\widehat{\alpha}_i)})$ with probability $1-\delta$. On the off diagonal entries, $\mathbb{E}[G] = \Pi^\top P\Pi$, and conditioned on $\Pi$, each entry of $G$ is an independent Bernoulli variable, therefore the total variance is bounded by $O(\log n/n\widehat{\alpha}_{\min})$. By Bernstein's inequality with high probability the deviation is bounded by $O(\log(nk/\delta)/\sqrt{n\widehat{\alpha}_{\min}})$ with probability $1-\delta$. For diagonal entries, $G_{i,i} = 0$, and $(\Pi^\top P\Pi)_{i,i} \le 1$. So the error in diagonal entries is at most the inner product of two rows in $Q$, which is equal to $O(\log n/n\widehat{\alpha}_{\min})$ (note that the diagonal entries do not dominate the error).

In the last step, we replace $Q$ by its estimate $\hat{Q}$. We shall write $\hat{Q}^i$ as $Q^i + \Delta_i$, where $\Delta_i$ is the $\ell_1$ norm perturbation of row $i$ then

$$\begin{aligned}(\hat{Q}G\hat{Q}^\top)_{i,j} - (QGQ^\top)_{i,j} &= (Q^i+\Delta_i)G(Q^j+\Delta_j)^\top - Q^iG(Q^j)^\top\\ &= \Delta_iG(Q^j+\Delta_j)^\top + Q^iG\Delta_j^\top\\ &\le O(|\Delta_i|_1 + |\Delta_j|_1)\end{aligned}$$

We now bound $|\Delta_i|_1$. With probability $1-2\delta$

$$\begin{aligned}|\Delta_i|_1 = |\hat{Q}^i - Q^i|_1 &\le \frac{(\alpha_0+1)}{n}\left(\widehat{\alpha}_{\min}^{-1}\varepsilon_\pi\sqrt{n\eta} + |\Pi^i|_1\varepsilon_T\widehat{\alpha}_{\min}^{-1/2}\right)\\ &\le O\left(\frac{(\alpha_0+1)^{3/2}\varepsilon_\pi}{\sqrt{n}}\widehat{\alpha}_{\min}^{-1}\widehat{\alpha}_{\max}^{1/2}\log\frac{nk}{\delta}\right),\end{aligned}$$

where the first term is due to $\ell_1$ guarantees on row-wise perturbation between $\hat{\Pi}$ and $\Pi$ in Lemma A.2 and the second term comes from eigenvalue perturbation in Lemma 4.2. We use the fact that since $\mathbb{E}[|\Pi^i|_1] = n\widehat{\alpha}_i$, $|\Pi^i|_1 \approx n\widehat{\alpha}_i$ with high probability. By expanding the expression we see that the last term dominates all the other terms. Therefore, the result holds. □

## A.3 Zero-error support recovery guarantees

Recall that we proposed Procedure 3 to provide improved support recovery estimates in the special case of homophilic models (where there are more edges within a community than to any community outside). We limit our analysis to the special case of uniform sized communities ($\alpha_i = 1/k$) and matrix $P$ such that $P(i,j) = p\mathbb{I}(i=j) + q\mathbb{I}(i \neq j)$ and $p \geq q$. In principle, the analysis can be extended to homophilic models with more general $P$ matrix (with suitably chosen thresholds for support recovery). We first consider analysis for the stochastic block model (i.e. $\lim \alpha_0 \to 0$).

*Proof of Corollary 4.5:* Since the threshold $\xi$ is $1/2$ in case of stochastic block models, we claim that the $\ell_1$ error for rows of $\hat{\Pi}$ is at most $O(\varepsilon_\pi^2)$ since $\Pi(i,j) \in \{0,1\}$, and in order for our method to make a mistake, it takes $1/4$ in the $\ell_2^2$ error. Thus, we view this as an approximation for community memberships (except for $O(\varepsilon_\pi^2)$ mistakes).

Notice that the next step (assigning communities for vertices in $A$) uses a completely different set of edges, therefore conditioned on $\Pi$, this step is independent with all previous steps. For each vertex we compute the average number of edges from this vertex to all the approximate communities, and set it to belong to the one with largest average degree. In order for this process to succeed, the error caused by the mistakes in approximate community (which is bounded by $(p-q)O(\varepsilon_\pi^2 \cdot k/|A|)$ because each mistake cause difference $p-q$ in the expected number of edges), must be smaller than the difference in the expected average number of edges (which is equal to $(p-q)$). At the same time, we need the average degree to constant well around its expectation, but this follows directly from Bernstein's inequality because the variance is bounded by $O(pk/n)$. Combining these requirements, we know the algorithm works when $\tilde{O}((p-q)\varepsilon_\pi^2 k/n) \leq (p-q)$, which implies

$$p - q \geq \tilde{\Omega}\left(\frac{\sqrt{p}k}{\sqrt{n}}\right).$$

□

We now prove the general result.

*Proof of Theorem 4.4:* We first show that the $\hat{F}_C$ computed by the algorithm is entry-wise $O((p-q)\varepsilon_P)$ close to the true $F_C$ matrix.

The proof is almost identical to the proof of A.3. The $\hat{Q}_B$ we have here is very similar to the matrix $\hat{Q}$ defined in A.3. Except that they have different support, the only difference is that we are making sure the $\ell_1$ norm of the first term is $(\alpha_0+1)$, while in A.3 it only has expected $\ell_1$ norm 1. Since the $\ell_1$ norm concentrates well this will be a low-order term in the final bound.

Similarly we define $Q_B^i = (\alpha_0+1)\frac{\Pi_B^i}{|\Pi_B^i|_1} + \frac{\alpha_0+1}{|B|}\vec{1}^\top$.

As in the proof of Lemma A.3 we have that $\Pi_B Q_B^\top$ is a matrix that is close to $I$, and

$$\begin{aligned}|(\Pi_C^\top P \Pi_B Q_B^\top)_{i,j} - (\Pi_C^\top P)_{i,j}| &\leq |(\Pi_C^\top P)^i((Q_B)_j - e_j)| = |(\Pi_C^\top P)^i \Delta_j| \\ &\leq O\left(\frac{pk \cdot \sqrt{\widehat{\alpha}_{max}/\widehat{\alpha}_{min}}}{\sqrt{n}}\sqrt{\log\frac{nk}{\delta}}\right).\end{aligned}$$

The reason this bound is factor $p$ smaller is because the entries in $(\Pi_C^\top P)^i$ are upperbounded by $p$ instead of 1. Like before this is actually a low order term.

Now we show if we replace $Q_B$ with $\hat{Q}_B$ here, the difference is also small using $|Q_B^i - \hat{Q}_B^i|_1 \leq O(\varepsilon_P)$

$$
\begin{aligned}
|(\Pi_C^\top P \Pi_B Q_B^\top)_{i,j} - (\Pi_C^\top P \Pi_B \hat{Q}_B^\top)_{i,j}| &= |(\Pi_C^\top P \Pi_B)^i (Q_B^\top - \hat{Q}_B^\top)_j| \\
&\le (\max \Pi_C^\top P \Pi_B - \min \Pi_C^\top P \Pi_B) |(Q_B^\top - \hat{Q}_B^\top)_j|_1 \\
&\le O((p-q)\varepsilon_P).
\end{aligned}
$$

Here we are using the fact that $(Q_B^j - \hat{Q}_B^j)\vec{1} = 0$, so we can subtract $\min \Pi_C^\top P \Pi_B$ from all the entries of $\Pi_C^\top P \Pi_B$ without influencing the result. That is also why we normalize the vectors $Q_B$ and $\hat{Q}_B$ (note that it is similar to computing the "average degree" instead of degree in the stochastic block model case).

Finally, $|(G_{C,B}\hat{Q}_B^\top)_{i,j}(\Pi_C^\top P \Pi_B \hat{Q}_B^\top)_{i,j}|$ are small by standard concentration bounds (and the differences are of lower order). Combining these we know $|\hat{F}_C(j,i) - F_C(j,i)| \le O((p-q)\varepsilon_P)$.

Similarly, we can show $|\hat{P}_{i,j} - P_{i,j}| \le O((p-q)\varepsilon_P)$ here. Which means the average of diagonals of $\hat{P}$ is $O((p-q)\varepsilon_P)$ close to $p$; the average of off-diagonals of $\hat{P}$ is $O((p-q)\varepsilon_P)$ close to $q$.

The entries of $\hat{F}_C$ are close to entries of $F_C$, and we know $\xi \gg \varepsilon_P$, therefore all the entries in $F_C$ that are larger than $q + (p-q)\xi$ must be found by our algorithm, and none of the entries that are smaller than $q + (p-q)\xi/2$ will not be called large by our algorithm. Finally the lemma follows because $\Pi_{i,j} = (F_{j,i} - q)/(p-q)$. $\square$

# B Tensor Power Method Analysis

In this section, we leverage on the perturbation analysis of Anandkumar et. al. [4]. However, we obtain stronger guarantees here through two modifications: (1)we modify the tensor deflation process in the robust power method in Procedure 2. Rather than a fixed deflation step after obtaining an estimate of the eigenvalue-eigenvector pair, in this paper, we deflate adaptively depending on the current estimate, and (2)rather than selecting random initialization vectors, as in [4], we initialize with vectors obtained from adjacency matrix. In Section B.1, we establish guarantees under "good" initialization vectors. This involves improved error bounds for the modified deflation procedure provided in Section B.2. In Section C.5, we establish that under Dirichlet distribution (for small $\alpha_0$), we obtain "good" initialization vectors.

## B.1 Analysis under good initialization vectors

We now show that when "good" initialization vectors are input to tensor power method in Procedure 2, we obtain good estimates of eigen-pairs under appropriate choice of number of iterations $N$ and spectral norm $\epsilon$ of tensor perturbation.

Let $T = \sum_{i\in[k]} \lambda_i v_i$, where $v_i$ are orthonormal vectors and $\lambda_1 \ge \lambda_2 \ge \dots \lambda_k$. Let $\tilde{T} = T + E$ be the perturbed tensor with $\|E\| \le \epsilon$. Recall that $N$ denotes the number of iteration of the tensor power method.

We call an initialization vector $u$ to be $(\gamma, R_0)$-good if there exists $v_i$ such that $\langle u, v_i\rangle > R_0$ and $|\langle u, v_i\rangle| - \max_{j<i} |\langle u, v_j\rangle| > \gamma |\langle u, v_i\rangle|$. Choose $\gamma = 1/100$.

**Theorem B.1.** *There exists universal constants $C_1, C_2 > 0$ such that the following holds.*

$$
\epsilon \le C_1 \cdot \lambda_{\min} R_0^2, \qquad N \ge C_2 \cdot \left(\log(k) + \log\log\left(\frac{\lambda_{\max}}{\epsilon}\right)\right), \tag{48}
$$

*Assume there is at least one good initialization vector corresponding to each $v_i$, $i \in [k]$. The parameter $\xi$ for choosing deflation vectors in each iteration of the tensor power method in Procedure 2 is chosen as $\xi \ge 25\epsilon$. We obtain eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{v}_1), (\hat{\lambda}_2, \hat{v}_2), \dots, (\hat{\lambda}_k, \hat{v}_k)$ such that there exists a permutation $\pi$ on $[k]$ with*

$$
\|v_{\pi(j)} - \hat{v}_j\| \le 8\epsilon/\lambda_{\pi(j)}, \qquad |\lambda_{\pi(j)} - \hat{\lambda}_j| \le 5\epsilon, \quad \forall j \in [k],
$$

*and*

$$\left\| T - \sum_{j=1}^{k} \hat{\lambda}_j \hat{v}_j^{\otimes 3} \right\| \leq 55\epsilon.$$

*Proof:* The proof is on lines of the proof of [4, Thm. 5.1] but here, we consider the modified deflation procedure, which improves the condition on $\epsilon$ in (48). We provide the full proof below for completeness.

We prove by induction on $i$, the number of eigenpairs estimated so far by Procedure 2. Assume that there exists a permutation $\pi$ on $[k]$ such that the following assertions hold.

1. For all $j \leq i$, $\|v_{\pi(j)} - \hat{v}_j\| \leq 8\epsilon/\lambda_{\pi(j)}$ and $|\lambda_{\pi(j)} - \hat{\lambda}_j| \leq 12\epsilon$.
2. $D(u,i)$ is the set of deflated vectors given current estimate of the power method is $u \in S^{k-1}$:
$$D(u,i;\xi) := \{j : |\hat{\lambda}_i \hat{\theta}_i| \geq \xi\} \cap [i],$$
where $\hat{\theta}_i := \langle u, \hat{v}_i \rangle$.
3. The error tensor
$$\tilde{E}_{i+1,u} := \Big(\hat{T} - \sum_{j \in D(u,i;\xi)} \hat{\lambda}_j \hat{v}_j^{\otimes 3}\Big) - \sum_{j \notin D(u,i;\xi)} \lambda_{\pi(j)} v_{\pi(j)}^{\otimes 3} = E + \sum_{j \in D(u,i;\xi)} \Big(\lambda_{\pi(j)} v_{\pi(j)}^{\otimes 3} - \hat{\lambda}_j \hat{v}_j^{\otimes 3}\Big)$$
satisfies
$$\|\tilde{E}_{i+1,u}(I,u,u)\| \leq 56\epsilon, \quad \forall u \in S^{k-1}; \tag{49}$$
$$\|\tilde{E}_{i+1,u}(I,u,u)\| \leq 2\epsilon, \quad \forall u \in S^{k-1} \text{ s.t. } \exists j \geq i+1 \,\centerdot\, (u^\top v_{\pi(j)})^2 \geq 1 - (168\epsilon/\lambda_{\pi(j)})^2. \tag{50}$$

We take $i = 0$ as the base case, so we can ignore the first assertion, and just observe that for $i = 0$, $D(u,0;\xi) = \emptyset$ and thus

$$\tilde{E}_{1,u} = \hat{T} - \sum_{j=1}^{k} \lambda_i v_i^{\otimes 3} = E, \quad \forall\, u \in S^{k-1}.$$

We have $\|\tilde{E}_1\| = \|E\| = \epsilon$, and therefore the second assertion holds.

Now fix some $i \in [k]$, and assume as the inductive hypothesis. The power iterations now take a subset of $j \in [i]$ for deflation, depending on the current estimate. Set

$$C_1 := \min\left\{(56 \cdot 9 \cdot 102)^{-1}, (100 \cdot 168)^{-1}, \Delta' \text{ from Lemma B.2 with } \Delta = 1/50\right\}. \tag{51}$$

For all good initialization vectors which are $\gamma$-separated relative to $\pi(j_{\max})$, we have (i) $|\theta^{(\tau)}_{j_{\max},0}| \geq R_0$, and (ii) that by in [4, Lemma B.4] (using $\tilde{\epsilon}/p := 2\epsilon$, $\kappa := 1$, and $i^* := \pi(j_{\max})$, and providing $C_2$),

$$|\tilde{T}_i(\theta_N^{(\tau)}, \theta_N^{(\tau)}, \theta_N^{(\tau)}) - \lambda_{\pi(j_{\max})}| \leq 5\epsilon$$

(notice by definition that $\gamma \geq 1/100$ implies $\gamma_0 \geq 1 - 1/(1+\gamma) \geq 1/101$, thus it follows from the bounds on the other quantities that $\tilde{\epsilon} = 2p\epsilon \leq 56 C_1 \cdot \lambda_{\min} R_0^2 < \frac{\gamma_0}{2(1+8\kappa)} \cdot \tilde{\lambda}_{\min} \cdot \theta^2_{i^*,0}$ as necessary). Therefore $\theta_N := \theta_N^{(\tau^*)}$ must satisfy

$$\tilde{T}_i(\theta_N, \theta_N, \theta_N) = \max_{\tau \in [L]} \tilde{T}_i(\theta_N^{(\tau)}, \theta_N^{(\tau)}, \theta_N^{(\tau)}) \geq \max_{j \geq i} \lambda_{\pi(j)} - 5\epsilon = \lambda_{\pi(j_{\max})} - 5\epsilon.$$

On the other hand, by the triangle inequality,

$$\begin{aligned}
\tilde{T}_i(\theta_N, \theta_N, \theta_N) &\leq \sum_{j \geq i} \lambda_{\pi(j)} \theta^3_{\pi(j),N} + |\tilde{E}_i(\theta_N, \theta_N, \theta_N)| \\
&\leq \sum_{j \geq i} \lambda_{\pi(j)} |\theta_{\pi(j),N}| \theta^2_{\pi(j),N} + 56\epsilon \\
&\leq \lambda_{\pi(j^*)} |\theta_{\pi(j^*),N}| + 56\epsilon
\end{aligned}$$

where $j^* := \arg\max_{j \geq i} \lambda_{\pi(j)} |\theta_{\pi(j),N}|$. Therefore

$$\lambda_{\pi(j^*)} |\theta_{\pi(j^*),N}| \geq \lambda_{\pi(j_{\max})} - 5\epsilon - 56\epsilon \geq \frac{4}{5} \lambda_{\pi(j_{\max})}.$$

Squaring both sides and using the fact that $\theta^2_{\pi(j^*),N} + \theta^2_{\pi(j),N} \leq 1$ for any $j \neq j^*$,

$$\begin{aligned} \left(\lambda_{\pi(j^*)} \theta_{\pi(j^*),N}\right)^2 &\geq \frac{16}{25} \left(\lambda_{\pi(j_{\max})} \theta_{\pi(j^*),N}\right)^2 + \frac{16}{25} \left(\lambda_{\pi(j_{\max})} \theta_{\pi(j),N}\right)^2 \\ &\geq \frac{16}{25} \left(\lambda_{\pi(j^*)} \theta_{\pi(j^*),N}\right)^2 + \frac{16}{25} \left(\lambda_{\pi(j)} \theta_{\pi(j),N}\right)^2 \end{aligned}$$

which in turn implies

$$\lambda_{\pi(j)} |\theta_{\pi(j),N}| \leq \frac{3}{4} \lambda_{\pi(j^*)} |\theta_{\pi(j^*),N}|, \quad j \neq j^*.$$

This means that $\theta_N$ is $(1/4)$-separated relative to $\pi(j^*)$. Also, observe that

$$|\theta_{\pi(j^*),N}| \geq \frac{4}{5} \cdot \frac{\lambda_{\pi(j_{\max})}}{\lambda_{\pi(j^*)}} \geq \frac{4}{5}, \quad \frac{\lambda_{\pi(j_{\max})}}{\lambda_{\pi(j^*)}} \leq \frac{5}{4}.$$

Therefore by [4, Lemma B.4] (using $\tilde{\epsilon}/p := 2\epsilon$, $\gamma := 1/4$, and $\kappa := 5/4$), executing another $N$ power iterations starting from $\theta_N$ gives a vector $\hat{\theta}$ that satisfies

$$\|\hat{\theta} - v_{\pi(j^*)}\| \leq \frac{8\epsilon}{\lambda_{\pi(j^*)}}, \qquad |\hat{\lambda} - \lambda_{\pi(j^*)}| \leq 5\epsilon.$$

Since $\hat{v}_i = \hat{\theta}$ and $\hat{\lambda}_i = \hat{\lambda}$, the first assertion of the inductive hypothesis is satisfied, as we can modify the permutation $\pi$ by swapping $\pi(i)$ and $\pi(j^*)$ without affecting the values of $\{\pi(j) : j \leq i-1\}$ (recall $j^* \geq i$).

We now argue that $\tilde{E}_{i+1,u}$ has the required properties to complete the inductive step. By Lemma B.2 (using $\tilde{\epsilon} := 5\epsilon$, $\xi = 5\tilde{\epsilon} = 25\epsilon$ and $\Delta := 1/50$, the latter providing one upper bound on $C_1$ as per (51)), we have for any unit vector $u \in S^{k-1}$,

$$\left\| \left( \sum_{j \leq i} \left( \lambda_{\pi(j)} v_{\pi(j)}^{\otimes 3} - \hat{\lambda}_j \hat{v}_j^{\otimes 3} \right) \right) (I, u, u) \right\| \leq \left( 1/50 + 100 \sum_{j=1}^{i} (u^\top v_{\pi(j)})^2 \right)^{1/2} 5\epsilon \leq 55\epsilon. \tag{52}$$

Therefore by the triangle inequality,

$$\|\tilde{E}_{i+1}(I, u, u)\| \leq \|E(I, u, u)\| + \left\| \left( \sum_{j \leq i} \left( \lambda_{\pi(j)} v_{\pi(j)}^{\otimes 3} - \hat{\lambda}_j \hat{v}_j^{\otimes 3} \right) \right) (I, u, u) \right\| \leq 56\epsilon.$$

Thus the bound (49) holds.

To prove that (50) holds, for any unit vector $u \in S^{k-1}$ such that there exists $j' \geq i+1$ with $(u^\top v_{\pi(j')})^2 \geq 1 - (168\epsilon/\lambda_{\pi(j')})^2$. We have (via the second bound on $C_1$ in (51) and the corresponding assumed bound $\epsilon \leq C_1 \cdot \lambda_{\min} R_0^2$)

$$100 \sum_{j=1}^{i} (u^\top v_{\pi(j)})^2 \leq 100 \left( 1 - (u^\top v_{\pi(j')})^2 \right) \leq 100 \left( \frac{168\epsilon}{\lambda_{\pi(j')}} \right)^2 \leq \frac{1}{50},$$

and therefore

$$\left( 1/50 + 100 \sum_{j=1}^{i} (u^\top v_{\pi(j)})^2 \right)^{1/2} 5\epsilon \leq (1/50 + 1/50)^{1/2} 5\epsilon \leq \epsilon.$$

By the triangle inequality, we have $\|\tilde{E}_{i+1}(I, u, u)\| \leq 2\epsilon$. Therefore (50) holds, so the second assertion of the inductive hypothesis holds. We conclude that by the induction principle, there exists a permutation $\pi$ such that two assertions hold for $i = k$. From the last induction step ($i = k$), it is also clear from (52) that $\|T - \sum_{j=1}^{k} \hat{\lambda}_j \hat{v}_j^{\otimes 3}\| \leq 55\epsilon$. This completes the proof of the theorem. □

## B.2 Deflation Analysis

**Lemma B.2** (Deflation analysis)**.** *Let $\tilde{\epsilon} > 0$ and let $\{v_1, \ldots, v_k\}$ be an orthonormal basis for $\mathbb{R}^k$ and $\lambda_i \geq 0$ for $i \in [k]$. Let $\{\hat{v}_1, \ldots, \hat{v}_k\} \in \mathbb{R}^k$ be a set of unit vectors and $\hat{\lambda}_i \geq 0$. Define third order tensor $\mathcal{E}_i$ such that*

$$\mathcal{E}_i := \lambda_i v_i^{\otimes 3} - \hat{\lambda}_i \hat{v}_i^{\otimes 3}, \quad \forall\, i \in k.$$

*For some $t \in [k]$ and a unit vector $u \in S^{k-1}$ such that $u = \sum_{i \in [k]} \theta_i v_i$ and $\hat{\theta}_i := \langle u, \hat{v}_i \rangle$, we have for $i \in [t]$,*

$$\begin{aligned} |\hat{\lambda}_i \hat{\theta}_i| &\geq \xi \geq 5\tilde{\epsilon}, \\ |\hat{\lambda}_i - \lambda_i| &\leq \tilde{\epsilon}, \\ \|\hat{v}_i - v_i\| &\leq \min\{\sqrt{2},\ 2\tilde{\epsilon}/\lambda_i\}, \end{aligned}$$

*then, the following holds*

$$\begin{aligned} \left\| \sum_{i=1}^{t} \mathcal{E}_i(I, u, u) \right\|_2^2 \leq \bigg( 4(5 + 11\tilde{\epsilon}/\lambda_{\min})^2 + 128(1 + \tilde{\epsilon}/\lambda_{\min})^2 (\tilde{\epsilon}/\lambda_{\min})^2 \bigg) \tilde{\epsilon}^2 \sum_{i=1}^{t} \theta_i^2 \\ + 64(1 + \tilde{\epsilon}/\lambda_{\min})^2 \tilde{\epsilon}^2 + 2048(1 + \tilde{\epsilon}/\lambda_{\min})^2 \tilde{\epsilon}^2. \end{aligned}$$

*In particular, for any $\Delta \in (0, 1)$, there exists a constant $\Delta' > 0$ (depending only on $\Delta$) such that $\tilde{\epsilon} \leq \Delta' \lambda_{\min}$ implies*

$$\left\| \sum_{i=1}^{t} \mathcal{E}_i(I, u, u) \right\|_2^2 \leq \bigg( \Delta + 100 \sum_{i=1}^{t} \theta_i^2 \bigg) \tilde{\epsilon}^2.$$

*Proof:* The proof is on lines of deflation analysis in [4, Lemma B.5], but we improve the bounds based on additional properties of vector $u$. From [4], we have that for all $i \in [t]$, and any unit vector $u$,

$$\begin{aligned} \left\| \sum_{i=1}^{t} \mathcal{E}_i(I, u, u) \right\|_2^2 \leq & \bigg( 4(5 + 11\tilde{\epsilon}/\lambda_{\min})^2 + 128(1 + \tilde{\epsilon}/\lambda_{\min})^2 (\tilde{\epsilon}/\lambda_{\min})^2 \bigg) \tilde{\epsilon}^2 \sum_{i=1}^{t} \theta_i^2 \\ & + 64(1 + \tilde{\epsilon}/\lambda_{\min})^2 \tilde{\epsilon}^2 \sum_{i=1}^{t} (\tilde{\epsilon}/\lambda_i)^2 + 2048(1 + \tilde{\epsilon}/\lambda_{\min})^2 \tilde{\epsilon}^2 \bigg( \sum_{i=1}^{t} (\tilde{\epsilon}/\lambda_i)^3 \bigg)^2. \end{aligned} \tag{53}$$

Let $\hat{\lambda}_i = \lambda_i + \delta_i$ and $\hat{\theta}_i = \theta_i + \beta_i$. We have $\delta_i \leq \tilde{\epsilon}$ and $\beta_i \leq 2\tilde{\epsilon}/\lambda_i$, and that $|\hat{\lambda}_i \hat{\theta}_i| \geq \xi$.

$$\begin{aligned} ||\hat{\lambda}_i \hat{\theta}_i| - |\lambda_i \theta_i|| &\leq |\hat{\lambda}_i \hat{\theta}_i - \lambda_i \theta_i| \\ &\leq |(\lambda_i + \delta_i)(\theta_i + \beta_i) - \lambda_i \theta_i| \\ &\leq |\delta_i \theta_i + \lambda_i \beta_i + \delta_i \beta_i| \\ &\leq 4\tilde{\epsilon}. \end{aligned}$$

Thus, we have that $|\lambda_i \theta_i| \geq 5\tilde{\epsilon} - 4\tilde{\epsilon} = \tilde{\epsilon}$. Thus $\sum_{i=1}^{t} \tilde{\epsilon}^2/\lambda_i^2 \leq \sum_i \theta_i^2 \leq 1$. Substituting in (53), we have the result. □

# C Concentration Bounds

## C.1 Main Result: Third Moment Tensor Perturbation Bound

**Notation:** Let $\|T\|$ denote the spectral norm for a tensor $T$ (or in special cases a matrix or a vector). Let $\|M\|_F$ denote the Frobenius norm. Let $|M_1|$ denote the operator $\ell_1$ norm, i.e., the maximum $\ell_1$ norm of

its columns and $\|M\|_\infty$ denote the maximum $\ell_1$ norm of its rows. Let $\kappa(M)$ denote the condition number, i.e., $\frac{\|M\|}{\sigma_{\min}(M)}$.

Unless otherwise specified, throughout the statements made mean that they occur with probability $1-\delta$ for a sufficiently small $\delta > 0$. Moreover, the probability space is the product space of the node communities, drawn from the Dirichlet distribution, and the edge variables drawn from the Bernoulli distribution $\mathrm{Ber}(P_{i,j})$ given the communities are $i$ and $j$ for the two edge points. Let $\tilde{O}$ denote $O(\cdot)$ up to poly log factors. We write $\Pi \sim \mathrm{Dir}(\alpha)$ to mean $\pi_i \overset{iid}{\sim} \mathrm{Dir}(\alpha)$, for $i \in V$.

We now provide the main result that the third-order whitened tensor computed from samples concentrates. Recall that $\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}$ denotes the third order moment computed using edges from partition $Y$ to partitions $A, B, C$ in (12). $\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC}$ are the whitening matrices defined in (21). The corresponding whitening matrices $W_A, W_B R_{AB}, W_C R_{AC}$ for exact moment third order tensor $\mathbb{E}[\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}|\Pi]$ will be defined later. Recall that $\rho$ is defined in (22) as

$$\rho := \frac{\alpha_0+1}{\widehat{\alpha}_{\min}}.$$

Given $\delta \in (0,1)$, throughout assume that

$$n = \Omega\left(\rho^2 \log^2 \frac{k}{\delta}\right). \tag{54}$$

**Theorem C.1** (Perturbation of whitened tensor)**.** *When the partitions $A, B, C, X, Y$ satisfy* (54)*, we have with probability* $1-100\delta$*,*

$$\begin{aligned}
\varepsilon_T &:= \left\| \mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC}) - \mathbb{E}[\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(W_A, \tilde{W}_B, \tilde{W}_C)|\Pi_A, \Pi_B, \Pi_C] \right\| \\
&= O\left( \frac{(\alpha_0+1)\sqrt{(\max_i(P\widehat{\alpha})_i)}}{n^{1/2}\widehat{\alpha}_{\min}^{3/2}\sigma_{\min}(P)} \cdot \left(1 + \left(\frac{\rho^2}{n}\log^2\frac{k}{\delta}\right)^{1/4}\right)\sqrt{\frac{\log k}{\delta}} \right) \\
&= \tilde{O}\left(\frac{\rho}{\sqrt{n}} \cdot \frac{\zeta}{\widehat{\alpha}_{\max}^{1/2}}\right),
\end{aligned} \tag{55}$$

## C.2 Whitening Matrix Perturbations

Consider rank-$k$ SVD of $|X|^{-1/2}(G^{\alpha_0}_{X,A})^\top_{k-svd} = \hat{U}_A\hat{D}_A\hat{V}_A^\top$, and the whitening matrix is given by $\hat{W}_A := \hat{U}_A\hat{D}_A^{-1}$ and thus $|X|^{-1}\hat{W}_A^\top(G^{\alpha_0}_{X,A})^\top(G^{\alpha_0}_{X,A})\hat{W}_A = I$. Now consider the singular value decomposition of

$$|X|^{-1}\hat{W}_A^\top\mathbb{E}[(G^{\alpha_0}_{X,A})^\top|\Pi] \cdot \mathbb{E}[(G^{\alpha_0}_{X,A})|\Pi]\hat{W}_A = \Phi\tilde{D}\Phi^\top.$$

$\hat{W}_A$ does not whiten the exact moments in general. On the other hand, consider

$$W_A := \hat{W}_A\Phi_A\tilde{D}_A^{-1/2}\Phi_A^\top.$$

Observe that $W_A$ whitens $|X|^{-1/2}\mathbb{E}[(G^{\alpha_0}_{X,A})|\Pi]$

$$|X|^{-1}W_A^\top\mathbb{E}[(G^{\alpha_0}_{X,A})^\top|\Pi]\mathbb{E}[(G^{\alpha_0}_{X,A})|\Pi]W_A = (\Phi_A\tilde{D}_A^{-1/2}\Phi_A^\top)^\top\Phi_A\tilde{D}_A\Phi_A^\top\Phi_A\tilde{D}_A^{-1/2}\Phi_A^\top = I$$

Now the ranges of $W_A$ and $\hat{W}_A$ may differ and we control the perturbations below.

Also note that $\hat{R}_{A,B}$, $\hat{R}_{A,C}$ are given by

$$\hat{R}_{AB} := |X|^{-1}\hat{W}_B^\top(G^{\alpha_0}_{X,B})^\top_{k-svd}(G^{\alpha_0}_{X,A})_{k-svd}\hat{W}_A. \tag{56}$$

$$R_{AB} := |X|^{-1}W_B^\top\mathbb{E}[(G^{\alpha_0}_{X,B})^\top|\Pi] \cdot \mathbb{E}[G^{\alpha_0}_{X,A}|\Pi] \cdot W_A. \tag{57}$$

Recall $\epsilon_G$ is given by (61), and $\sigma_{\min}\left(\mathbb{E}[G^{\alpha_0}_{X,A}|\Pi]\right)$ is given in (C.8) and $|A| = |B| = |X| = n$.

**Lemma C.2** (Whitening matrix perturbations). *With probability* $1-\delta$*,*

$$\epsilon_{W_A} := \| \operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top (\hat{W}_A - W_A)\| = O\left(\frac{(1-\varepsilon_1)^{-1/2}\epsilon_G}{\sigma_{\min}\left(\mathbb{E}[G_{X,A}^{\alpha_0}|\Pi]\right)}\right) \tag{58}$$

$$\epsilon_{\tilde{W}_B} := \| \operatorname{Diag}(\widehat{\alpha})^{1/2} F_B^\top (\hat{W}_B \hat{R}_{AB} - W_B R_{AB})\| = O\left(\frac{(1-\varepsilon_1)^{-1/2}\epsilon_G}{\sigma_{\min}\left(\mathbb{E}[G_{X,B}^{\alpha_0}|\Pi]\right)}\right) \tag{59}$$

*Thus, with probability* $1-6\delta$*,*

$$\epsilon_{W_A} = \epsilon_{\tilde{W}_B} = O\left(\frac{(\alpha_0+1)\sqrt{\max_i (P\widehat{\alpha})_i}}{n^{1/2}\widehat{\alpha}_{\min}\sigma_{\min}(P)} \cdot (1+\varepsilon_1+\varepsilon_2+\varepsilon_3)\right), \tag{60}$$

*where* $\varepsilon_1$*,* $\varepsilon_2$ *and* $\varepsilon_3$ *are given by* (67) *and* (68)*.*

**Remark:** Note that when partitions $X, A$ satisfy (54), $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are small. When $P$ is well conditioned and $\widehat{\alpha}_{\min} = \widehat{\alpha}_{\max} = 1/k$, we have $\epsilon_{W_A}, \epsilon_{\tilde{W}_B} = O(k/\sqrt{n})$.

*Proof:* Using the fact that $W_A = \hat{W}_A \Phi_A \tilde{D}_A^{-1/2} \Phi_A^\top$ or $\hat{W}_A = W_A \Phi_A \tilde{D}_A^{1/2} \Phi_A^\top$ we have that

$$\begin{aligned}
\| \operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top (\hat{W}_A - W_A)\| &\leq \| \operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top W_A (I - \Phi_A \tilde{D}_A^{1/2}\Phi_A^\top)\| \\
&= \| \operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top W_A (I - \tilde{D}_A^{1/2})\| \\
&\leq \| \operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top W_A (I - \tilde{D}_A^{1/2})(I + \tilde{D}_A^{1/2})\| \\
&\leq \| \operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top W_A\| \cdot \|I - \tilde{D}_A\|
\end{aligned}$$

using the fact that $\tilde{D}_A$ is a diagonal matrix.

Now note that $W_A$ whitens $|X|^{-1/2}\mathbb{E}[G_{X,A}^{\alpha_0}|\Pi] = |X|^{-1/2} F_A \operatorname{Diag}(\alpha^{1/2})\Psi_X$, where $\Psi_X$ is defined in (66). Further it is shown in Lemma C.8 that $\Psi_X$ satisfies with probability $1-\delta$ that

$$\varepsilon_1 := \|I - |X|^{-1}\Psi_X \Psi_X^\top\| \leq O\left(\sqrt{\frac{(\alpha_0+1)}{\widehat{\alpha}_{\min}|X|}} \cdot \log\frac{k}{\delta}\right)$$

Since $\varepsilon_1 \ll 1$ when $X, A$ satisfy (54). We have that $|X|^{-1/2}\Psi_X$ has singular values around 1. Since $W_A$ whitens $|X|^{-1/2}\mathbb{E}[G_{X,A}^{\alpha_0}|\Pi]$, we have

$$|X|^{-1} W_A^\top F_A \operatorname{Diag}(\alpha^{1/2})\Psi_X \Psi_X^\top \operatorname{Diag}(\alpha^{1/2}) F_A^\top W_A = I.$$

Thus, with probability $1-\delta$,

$$\| \operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top W_A\| = O((1-\varepsilon_1)^{-1/2}).$$

Let $\mathbb{E}[(G_{X,A}^{\alpha_0})|\Pi] = (G_{X,A}^{\alpha_0})_{k-svd} + \Delta$. We have

$$\begin{aligned}
\|I - \tilde{D}_A\| &= \|I - \Phi_A \tilde{D}_A \Phi_A^\top\| \\
&= \|I - |X|^{-1}\hat{W}_A^\top \mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi] \cdot \mathbb{E}[(G_{X,A}^{\alpha_0})|\Pi]\hat{W}_A\| \\
&= O\left(|X|^{-1}\|\hat{W}_A^\top \left(\Delta^\top (G_{X,A}^{\alpha_0})_{k-svd} + \Delta (G_{X,A}^{\alpha_0})_{k-svd}^\top\right)\hat{W}_A\|\right) \\
&= O\left(|X|^{-1/2}\|\hat{W}_A^\top \Delta^\top \hat{V}_A + \hat{V}_A^\top \Delta \hat{W}_A\|\right), \\
&= O\left(|X|^{-1/2}\|\hat{W}_A\|\|\Delta\|\right) \\
&= O\left(|X|^{-1/2}\|W_A\|\epsilon_G\right),
\end{aligned}$$

since $\|\Delta\| \leq \epsilon_G + \sigma_{k+1}(G^{\alpha_0}_{X,A}) \leq 2\epsilon_G$, using Weyl's theorem for singular value perturbation and the fact that $\epsilon_G \cdot \|W_A\| \ll 1$ and $\|W_A\| = |X|^{1/2}/\sigma_{\min}\left(\mathbb{E}[G^{\alpha_0}_{X,A}|\Pi]\right)$.

We now consider perturbation of $W_B R_{AB}$. By definition, we have that

$$\mathbb{E}[G^{\alpha_0}_{X,B}|\Pi] \cdot W_B R_{AB} = \mathbb{E}[G^{\alpha_0}_{X,A}|\Pi] \cdot W_A.$$

and

$$\|W_B R_{AB}\| = |X|^{1/2}\sigma_{\min}(\mathbb{E}[G^{\alpha_0}_{X,B}|\Pi])^{-1}.$$

Along the lines of previous derivation for $\epsilon_{W_A}$, let

$$|X|^{-1}(\hat{W}_B\hat{R}_{AB})^\top \cdot \mathbb{E}[(G^{\alpha_0}_{X,B})^\top|\Pi] \cdot \mathbb{E}[G^{\alpha_0}_{X,B}|\Pi]\hat{W}_B\hat{R}_{AB} = \Phi_B\tilde{D}_B\Phi_B^\top.$$

Again using the fact that $|X|^{-1}\Psi_X\Psi_X^\top \approx I$, we have

$$\|\operatorname{Diag}(\widehat{\alpha})^{1/2}F_B^\top W_B R_{AB}\| \approx \|\operatorname{Diag}(\widehat{\alpha})^{1/2}F_A^\top W_A\|,$$

and the rest of the proof follows. □

## C.3 Proving the tensor concentration bound

*Proof of Theorem C.1:* In tensor $T^{\alpha_0}$ in (12), the first term is

$$(\alpha_0+1)(\alpha_0+2)\sum_{i\in Y}\left(G_{i,A}^\top \otimes G_{i,B}^\top \otimes G_{i,C}^\top\right).$$

We claim that this term dominates in the perturbation analysis since the mean vector perturbation is of lower order. We now consider perturbation of the whitened tensor

$$\Lambda_0 = \frac{1}{|Y|}\sum_{i\in Y}\left((\hat{W}_A^\top G_{i,A}^\top) \otimes (\hat{R}_{AB}^\top\hat{W}_B^\top G_{i,B}^\top) \otimes (\hat{R}_{AC}^\top\hat{W}_C^\top G_{i,C}^\top)\right).$$

We show that this tensor is close to the corresponding term in the expectation in three steps.

First we show it is close to

$$\Lambda_1 = \frac{1}{|Y|}\sum_{i\in Y}\left((\hat{W}_A^\top F_A\pi_i) \otimes (\hat{R}_{AB}^\top\hat{W}_B^\top F_B\pi_i) \otimes (\hat{R}_{AC}^\top\hat{W}_C^\top F_C\pi_i)\right).$$

Then this vector is close to the expectation over $\Pi_Y$.

$$\Lambda_2 = \mathbb{E}_{\pi\sim\operatorname{Dir}(\alpha)}\left((\hat{W}_A^\top F_A\pi) \otimes (\hat{R}_{AB}^\top\hat{W}_B^\top F_B\pi) \otimes (\hat{R}_{AC}^\top\hat{W}_C^\top F_C\pi)\right).$$

Finally we replace the estimated whitening matrix $\hat{W}_A$ with $W_A$.

$$\Lambda_3 = \mathbb{E}_{\pi\sim\operatorname{Dir}(\alpha)}\left((W_A^\top F_A\pi) \otimes (\tilde{W}_B^\top F_B\pi) \otimes (\tilde{W}_C^\top F_C\pi)\right).$$

For $\Lambda_0 - \Lambda_1$, the dominant term in the perturbation bound (assuming partitions $A, B, C, X, Y$ are of size $n$) is (since for any rank 1 tensor, $\|u \otimes v \otimes w\| = \|u\| \cdot \|v\| \cdot \|w\|$),

$$O\left(\frac{1}{|Y|}\|\tilde{W}_B^\top F_B\|^2\left\|\sum_{i\in Y}\left(\hat{W}_A^\top G_{i,A}^\top - \hat{W}_A^\top F_A\pi_i\right)\right\|\right)$$
$$O\left(\frac{1}{|Y|}\widehat{\alpha}_{\min}^{-1}\cdot\frac{(\alpha_0+1)(\max_i(P\widehat{\alpha})_i)}{\widehat{\alpha}_{\min}\sigma_{\min}(P)}\cdot(1+\varepsilon_1+\varepsilon_2+\varepsilon_3)\sqrt{\log\frac{n}{\delta}}\right),$$

with probability $1-13\delta$ (Lemma C.3). Since there are 7 terms in the third order tensor $\mathrm{T}^{\alpha_0}$, we have the bound with probability $1-91\delta$.

For $\Lambda_1-\Lambda_2$, since $\hat{W}_A F_A \operatorname{Diag}(\widehat{\alpha})^{1/2}$ has spectral norm almost 1, by Lemma C.5 the spectral norm of the perturbation is at most

$$\begin{aligned}&\left\|\hat{W}_A F_A \operatorname{Diag}(\widehat{\alpha})^{1/2}\right\|^3 \left\|\frac{1}{|Y|}\sum_{i\in Y}(\operatorname{Diag}(\widehat{\alpha})^{-1/2}\pi_i)^{\otimes 3} - \mathbb{E}_{\pi\sim\operatorname{Dir}(\alpha)}(\operatorname{Diag}(\widehat{\alpha})^{-1/2}\pi_i)^{\otimes 3}\right\| \\ &\leq O(1/\widehat{\alpha}_{min}\sqrt{n}\sqrt{\log n/\delta}).\end{aligned}$$

For the final term $\Lambda_2-\Lambda_3$, the dominating term is

$$(\hat{W}_A - W_A)F_A \operatorname{Diag}(\widehat{\alpha})^{1/2}\left\|\Lambda_3\right\| \leq \varepsilon_{W_A}\left\|\Lambda_3\right\| \leq O\left(\frac{(\alpha_0+1)\sqrt{\max_i(P\widehat{\alpha})_i}}{n^{1/2}\widehat{\alpha}_{\min}^{3/2}\sigma_{\min}(P)}(1+\varepsilon_1+\varepsilon_2+\varepsilon_3)\sqrt{\log\frac{n}{\delta}}\right)$$

Putting all these together, the third term $\|\Lambda_2-\Lambda_3\|$ dominates. We know with probability at least $1-100\delta$, the perturbation in the tensor is at most

$$O\left(\frac{(\alpha_0+1)\sqrt{\max_i(P\widehat{\alpha})_i}}{n^{1/2}\widehat{\alpha}_{\min}^{3/2}\sigma_{\min}(P)}(1+\varepsilon_1+\varepsilon_2+\varepsilon_3)\sqrt{\log\frac{n}{\delta}}\right).$$

□

**Lemma C.3** (Concentration of sum of whitened vectors)**.** *Assuming all the partitions satisfy* (54)*, with probability* $1-7\delta$*,*

$$\begin{aligned}\left\|\sum_{i\in Y}\left(\hat{W}_A^\top G_{i,A}^\top - \hat{W}_A^\top F_A\pi_i\right)\right\| &= O(\sqrt{|Y|\widehat{\alpha}_{\max}}\epsilon_{W_A}) \\ &= O\left(\frac{\sqrt{(\alpha_0+1)(\max_i(P\widehat{\alpha})_i)}}{\widehat{\alpha}_{\min}\sigma_{\min}(P)}\cdot(1+\varepsilon_2+\varepsilon_3)\sqrt{\log n/\delta}\right), \\ \left\|\sum_{i\in Y}\left((\hat{W}_B\hat{R}_{AB})^\top(G_{i,B}^\top - F_B\pi_i)\right)\right\| &= O\left(\frac{\sqrt{(\alpha_0+1)(\max_i(P\widehat{\alpha})_i)}}{\widehat{\alpha}_{\min}\sigma_{\min}(P)}\cdot(1+\varepsilon_1+\varepsilon_2+\varepsilon_3)\sqrt{\log n/\delta}\right).\end{aligned}$$

**Remark:** Note that when $P$ is well conditioned and $\widehat{\alpha}_{\min}=\widehat{\alpha}_{\max}=1/k$, we have the above bounds as $O(k)$. Thus, when it is normalized with $1/|Y|=1/n$, we have the bound as $O(k/n)$.

*Proof:* Note that $\hat{W}_A$ is computed using partition $X$ and $G_{i,A}$ is obtained from $i\in Y$. We have independence for edges across different partitions $X$ and $Y$. Let $\Xi_i := \hat{W}_A^\top(G_{i,A}^\top - F_A\pi_i)$.Applying matrix Bernstein's inequality to each of the variables, we have

$$\begin{aligned}\|\Xi_i\| &\leq \|\hat{W}_A\|\cdot\|G_{i,A}^\top - F_A\pi_i\| \\ &\leq \|\hat{W}_A\|\sqrt{\|F_A\|_1},\end{aligned}$$

from Lemma C.7. The variances are given by

$$\begin{aligned}\|\sum_{i\in Y}\mathbb{E}[\Xi_i\Xi_i^\top|\Pi]\| &\leq \sum_{i\in Y}\hat{W}_A^\top \operatorname{Diag}(F_A\pi_i)\hat{W}_A, \\ &\leq \|\hat{W}_A\|^2\|F_Y\|_1 \\ &= O\left(\frac{|Y|}{|A|}\cdot\frac{(\alpha_0+1)(\max_i(P\widehat{\alpha})_i)}{\widehat{\alpha}_{\min}^2\sigma_{\min}^2(P)}\cdot(1+\varepsilon_2+\varepsilon_3)\right),\end{aligned}$$

with probability $1-2\delta$ from (64) and (65), and $\varepsilon_2, \varepsilon_3$ are given by (68). Similarly, $\|\sum_{i\in Y}\mathbb{E}[\Xi_i^\top \Xi_i|\Pi]\| \leq \|\hat{W}_A\|^2\|F_Y\|_1$. Thus, from matrix Bernstein's inequality, we have with probability $1-3\delta$

$$\|\sum_{i\in Y}\Xi_i\| = O(\|\hat{W}_A\|\sqrt{\max(\|F_A\|_1, \|F_X\|_1)}).$$

$$= O\left(\frac{\sqrt{(\alpha_0+1)(\max_i(P\widehat{\alpha})_i)}}{\widehat{\alpha}_{\min}\sigma_{\min}(P)}\cdot(1+\varepsilon_2+\varepsilon_3)\sqrt{\log n/\delta}\right)$$

On similar lines, we have the result for $B$ and $C$, and also use the independence assumption on edges in various partitions. □

We now show that not only the sum of whitened vectors concentrates, but that each individual whitened vector $\hat{W}_A^\top G_{i,A}^\top$ concentrates, when $A$ is large enough.

**Lemma C.4** (Concentration of a random whitened vector)**.** *Conditioned on* $\Pi$ *matrix, with probability at least* $1/4$*,*

$$\left\|\hat{W}_A^T G_{i,A}^T - W_A^T F_A\pi_i\right\| \leq O(\varepsilon_{W_A}\widehat{\alpha}_{min}^{-1/2}) = \tilde{O}\left(\frac{\sqrt{(\alpha_0+1)(\max_i(P\widehat{\alpha})_i)}}{n^{1/2}\widehat{\alpha}_{\min}^{3/2}\sigma_{\min}(P)}\right).$$

*Proof.* We have

$$\left\|\hat{W}_A^T G_{i,A}^T - W_A^T F_A\pi_i\right\| \leq \left\|(\hat{W}_A - W_A)^T F_A\pi_i\right\| + \left\|\hat{W}_A^T(G_{i,A}^T - \hat{W}_A^T F_A\pi)\right\|.$$

The first term is satisfies satisfies with probability $1-3\delta$

$$\|(\hat{W}_A^\top - W_A^\top)F_A\pi_i\| \leq \epsilon_{W_A}\widehat{\alpha}_{min}^{-1/2}$$
$$O\left(\frac{(\alpha_0+1)\widehat{\alpha}_{\max}^{1/2}\sqrt{(\max_i(P\widehat{\alpha})_i)}}{n^{1/2}\widehat{\alpha}_{\min}^{3/2}\sigma_{\min}(P)}\cdot(1+\varepsilon_1+\varepsilon_2+\varepsilon_3)\right)$$

Now we bound the second term. Note that $G_{i,A}^\top$ is independent of $\hat{W}_A^\top$, since they are related to disjoint subset of edges. The whitened neighborhood vector can be viewed as a sum of vectors:

$$\hat{W}_A^\top G_{i,A}^\top = \sum_{j\in A} G_{i,j}(\hat{W}_A^\top)_j = \sum_{j\in A} G_{i,j}(\hat{D}_A\hat{U}_A^\top)_j = \hat{D}_A\sum_{j\in A} G_{i,j}(\hat{U}_A^\top)_j.$$

Conditioned on $\pi_i$ and $F_A$, $G_{i,j}$ are Bernoulli variables with probability $(F_A\pi_i)_j$. The goal is to compute the variance of the sum, $\sum_{j\in A}(F_A\pi_i)_j\left\|(\hat{U}_A^\top)_j\right\|^2$, and then use Chebyshev's inequality.

By Wedin's theorem, we know the span of columns of $\hat{U}_A$ is $O(\epsilon_G/\sigma_{min}(G_X^{\alpha_0}, A)) = O(\epsilon_{W_A})$ close to the span of columns of $F_A$. The span of columns of $F_A$ is the same as the span of rows in $\Pi_A$. In particular, let $Proj_\Pi$ be the projection matrix of the span of rows in $\Pi_A$, we have

$$\left\|\hat{U}_A\hat{U}_A^\top - Proj_\Pi\right\| \leq O(\epsilon_{W_A}).$$

Using the spectral norm bound, we have the Frobenius norm

$$\left\|\hat{U}_A\hat{U}_A^\top - Proj_\Pi\right\|_F \leq O(\epsilon_{W_A}\sqrt{k})$$

since they are rank $k$ matrices. This implies that

$$\sum_{j\in A}\left(\left\|(\hat{U}_A^T)_j\right\| - \left\|Proj_\Pi^j\right\|\right)^2 = O(\epsilon_{W_A}^2 k).$$

Now

$$\|Proj_\Pi^j\| \leq \frac{\|\pi_j\|}{\sigma_{\min}(\Pi_A)} = O\left(\sqrt{\frac{(\alpha_0+1)}{n\widehat{\alpha}_{\min}}}\right),$$

from Lemma C.8

Now we can bound the variance of the vectors $\sum_{j\in A} G_{i,j}(\hat{U}_A^T)_j$, since the variance of $G_{i,j}$ is bounded by $(F_A\pi_i)_j$ (its probability), and the variance of the vectors is at most

$$\begin{aligned}
\sum_{j\in A}(F_A\pi_i)_j \left\|(\hat{U}_A^\top)_j\right\|^2 &\leq 2\sum_{j\in A}(F_A\pi_i)_j \left\|Proj_\Pi^j\right\|^2 + 2\sum_{j\in A}(F_A\pi_i)_j \left(\left\|(\hat{U}_A^T)_j\right\| - \left\|Proj_\Pi^j\right\|\right)^2 \\
&\leq 2\sum_{j\in A}(F_A\pi_i)_j \max_{j\in A}\left(\left\|Proj_\Pi^j\right\|^2\right) + \max_{i,j} P_{i,j}\sum_{j\in A}\left(\left\|(\hat{U}_A^T)_j\right\| - \left\|Proj_\Pi^j\right\|\right)^2 \\
&\leq O\left(\frac{|F_A|_1(\alpha_0+1)}{n\widehat{\alpha}_{\min}}\right)
\end{aligned}$$

Now Chebyshev's inequality implies that with probability at least 1/4 (or any other constant),

$$\left\|\sum_{j\in A}(G_{i,j} - F_A\pi_i)(\hat{U}_A^T)_j\right\|^2 \leq O\left(\frac{|F_A|_1(\alpha_0+1)}{n\widehat{\alpha}_{\min}}\right).$$

And thus, we have

$$\hat{W}_A^T(G_{i,A} - F_A\pi_i) \leq \sqrt{\frac{|F_A|_1(\alpha_0+1)}{n\widehat{\alpha}_{\min}}}\cdot\left\|\hat{W}_A^T\right\| \leq O\left(\epsilon_{W_A}\widehat{\alpha}_{\min}^{-1/2}\right).$$

Combining the two terms, we have the result. ∎

**Remark:** This lemma is used to show that with some probability $\hat{W}^T G_{i,A}$ is a good initial vector for tensor power method. In order to achieve that we need to combine this Lemma with Lemma C.10.

Notice that there are three events here: (1) $\epsilon_{W_A}$ is small and $|F_A|_1$ is concentrated around its expectation; (2) $\left\|\hat{W}^T G_{i,A} - \hat{W}^T F_A\pi_i\right\|$ is small; (3) $W_A^\top F_A\pi_i$ is a $(\gamma, r_0)$ good initial vector.

When all three events happen we can conclude $\hat{W}^\top G_{i,A}$ is $(\gamma - \frac{2\Delta}{r_0-\Delta}, r_0-\Delta)$ good. Event (1) happens with high probability. Event (3) is only related to properties of $\Pi$. The proof above shows conditioned on Event (1) and any value of $\Pi$, the probability that Event (2) happens for index $i$ is always at least 1/4, therefore the intersection of three events happens with significant probability.

**Lemma C.5.** *With probability* $1-\delta$*,*

$$\begin{aligned}
\left\|\frac{1}{|Y|}\sum_{i\in Y}(\mathrm{Diag}(\widehat{\alpha})^{-1/2}\pi_i)^{\otimes 3} - \mathbb{E}_{\pi\sim\mathrm{Dir}(\alpha)}(\mathrm{Diag}(\widehat{\alpha})^{-1/2}\pi_i)^{\otimes 3}\right\| &\leq O(\cdot\frac{1}{\widehat{\alpha}_{min}\sqrt{n}}\sqrt{\log n/\delta}) \\
&= \tilde{O}(\frac{1}{\widehat{\alpha}_{min}\sqrt{n}})
\end{aligned}$$

*Proof.* The spectral norm of this tensor cannot be larger than the spectral norm of a $k\times k^2$ matrix that we obtain be "collapsing" the last two dimensions (by definitions of norms). Let $\phi_i = \mathrm{Diag}(\hat{\alpha})^{-1/2}\pi_i$, the "collapsed" tensor is just the matrix $\phi_i(\phi_i\otimes\phi_i)^\top$ (here we view $\phi_i\otimes\phi_i$ as a vector in $\mathbb{R}^{k^2}$). We apply Matrix Bernstein on the matrices $Z_i = \phi_i(\phi_i\otimes\phi_i)^\top$.

Clearly, $\left\|\sum_{i\in Y}\mathbb{E}[Z_iZ_i^\top]\right\| \leq |Y|\max\|\phi\|^4\left\|\mathbb{E}[\phi\phi^\top]\right\| \leq |Y|\widehat{\alpha}_{min}^{-2}$ because $\left\|\mathbb{E}[\phi\phi^\top]\right\| \leq 2$.

For the other variance term $\left\|\sum_{i\in Y}\mathbb{E}[Z_i^\top Z_i]\right\|$, we have $\left\|\sum_{i\in Y}\mathbb{E}[Z_i^\top Z_i]\right\| \leq |Y|\widehat{\alpha}_{min}\left\|\mathbb{E}[(\phi\otimes\phi)(\phi\otimes\phi)^\top]\right\|$.

It remains to bound the norm of $\mathbb{E}[(\phi \otimes \phi)(\phi \otimes \phi)^\top]$. Suppose we have a variable $X = \sum_{i,j} A_{i,j}\phi(i)\phi(j)$, then by definition the spectral norm of $\mathbb{E}[(\phi \otimes \phi)(\phi \otimes \phi)^\top]$ is the maximum expectation of $X^2$ among all matrices $A$ such that $\|A\|_F = 1$.

We group the terms in $\mathbb{E}[X^2]$ according to the powers $\phi$ variables, and then bound the contribution of different groups of terms separately, and $\mathbb{E}[X^2]$ is at most the sum of absolute values of different groups.

$\phi(i)^4$ terms: By properties of Dirichlet distribution we know $\mathbb{E}[\phi(i)^4] = \Theta(\widehat{\alpha}_i^{-1}) \leq O(\widehat{\alpha}_{min}^{-1})$. The coefficients in front of $\phi(i)^4$ is $A_{i,i}^2$, so we know

$$|\mathbb{E}[A_{i,i}^2\phi(i)^4| \leq O(\widehat{\alpha}_{min}^{-1})\sum_{i=1}^{k} A_{i,i}^2 \leq O(\widehat{\alpha}_{min}^{-1}).$$

$\phi(i)^3\phi(j)$ terms: Ignore symmetries (there are constantly many so we lose at most a constant here) the coefficients are $A_{i,i}A_{i,j}$. The expectation of $\phi(i)^3\phi(j)$ is $\Theta(\hat{\alpha}(i)^{-1/2}\hat{\alpha}(j)^{1/2})$. Therefore, the total contribution of these terms is bounded by

$$|\mathbb{E}[\sum_{i,j} A_{i,i}A_{i,j}\phi(i)^3\phi(j)]| \leq O(\sqrt{\sum_{i,j}(A_{i,i}^2\hat{\alpha}(j))\sum_{i,j} A_{i,j}^2\hat{\alpha}(i)^{-1}}) \leq O(\widehat{\alpha}_{min}^{-1/2}).$$

$\phi(i)^2\phi(j)^2$ terms: the total number of such terms is $O(k^2)$, the expectation of $\phi(i)^2\phi(j)^2$ is $\Theta(1)$, so the Frobenius norm of that part of matrix is smaller than $O(k)$ (which implies $|\mathbb{E}[\sum_{i,j}(A_{i,j}^2+2A_{i,i}A_{j,j})\phi(i)^2\phi(j)^2]| \leq O(k)$).

$\phi(i_1)^2\phi(i_2)\phi(i_3)$ terms: similarly, there are $O(k^3)$ such terms, each one has expectation $\Theta(\hat{\alpha}(i_2)^{1/2}\hat{\alpha}(i_3)^{1/2})$. The Frobenius norm of this part of matrix is bounded by

$$O\left(\sqrt{\sum_{i_1,i_2,i_3\in[k]} \hat{\alpha}(i_2)\hat{\alpha}(i_3)}\right) \leq O(\sqrt{k})\sqrt{\sum_{i_2}\sum_{i_3}\widehat{\alpha}_{i_2}\widehat{\alpha}_{i_3}} \leq O(\sqrt{k}).$$

the rest: the sum is $\mathbb{E}[\sum_{i,j,p,q} A_{i,j}A_{p,q}\hat{\alpha}(i)^{1/2}\hat{\alpha}(j)^{1/2}\hat{\alpha}(p)^{1/2}\hat{\alpha}(q)^{1/2}]$. It is easy to break the bounds into the product of two sums ($\sum_{i,j}$ and $\sum_{p,q}$) and then bound each one by Cauchy-Schwartz, the result is 1.

Hence the variance term in Matrix Bernstein's inequality can be bounded by $\sigma^2 \leq O(n\widehat{\alpha}_{min}^{-2})$, each term has norm at most $\widehat{\alpha}_{min}^{-3/2}$. When $\widehat{\alpha}_{min}^{-2} < n$ we know the variance term dominates and the spectral norm of the difference is at most $O(\widehat{\alpha}_{min}^{-1}n^{-1/2}\sqrt{\log n/\delta})$ with probability $1-\delta$.

□

## C.4 Concentration of adjacency matrix and other auxiliary lemmata

Let $n := \max(|A|, |X|)$.

**Lemma C.6** (Concentration of $G_{X,A}^{\alpha_0}$)**.** *When $\pi_i \sim \mathrm{Dir}(\alpha)$, for $i \in V$, with probability $1-4\delta$,*

$$\epsilon_G := \|G_{X,A}^{\alpha_0} - \mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi]\| = O\left(\sqrt{(\alpha_0+1)n\cdot(\max_i(P\widehat{\alpha})_i)(1+\varepsilon_2)\log\frac{n}{\delta}}\right) \tag{61}$$

*Proof:* From definition of $G_{X,A}^{\alpha_0}$, we have

$$\epsilon_G \leq \sqrt{\alpha_0+1}\|G_{X,A} - \mathbb{E}[G_{X,A}|\Pi]\| + (\sqrt{\alpha_0+1}-1)\sqrt{|X|}\|\mu_{X,A} - \mathbb{E}[\mu_{X,A}|\Pi]\|.$$

We have concentration for $\mu_{X,A}$ and adjacency submatrix $G_{X,A}$ from Lemma C.7. □

We now provide concentration bounds for adjacency sub-matrix $G_{X,A}$ from partition $X$ to $A$ and the corresponding mean vector. Recall that $\mathbb{E}[\mu_{X\to A}|F_A,\pi_X] = F_A\pi_X$ and $\mathbb{E}[\mu_{X\to A}|F_A] = F_A\widehat{\alpha}$.

**Lemma C.7** (Concentration of adjacency submatrices)**.** *When $\pi_i \overset{iid}{\sim} \mathrm{Dir}(\alpha)$ for $i \in V$, with probability $1-2\delta$,*

$$\|G_{X,A} - \mathbb{E}[G_{X,A}|\Pi]\| = O\left(\sqrt{n \cdot (\max_i (P\widehat{\alpha})_i)(1+\varepsilon_2)\log\frac{n}{\delta}}\right). \tag{62}$$

$$\|\mu_A - \mathbb{E}[\mu_A|\Pi]\| = O\left(\frac{1}{|X|}\sqrt{n \cdot (\max_i (P\widehat{\alpha})_i)(1+\varepsilon_2)\log\frac{n}{\delta}}\right), \tag{63}$$

*where $\varepsilon_2$ is given by* (68)*.*

*Proof:* Recall $\mathbb{E}[G_{X,A}|\Pi] = F_A\Pi_X$ and $G_{A,X} = \mathrm{Ber}(F_A\Pi_X)$ where $\mathrm{Ber}(\cdot)$ denotes the Bernoulli random matrix with independent entries. Let

$$Z_i := (G_{i,A}^\top - F_A\pi_i)e_i^\top.$$

We have $G_{X,A}^\top - F_A\Pi_X = \sum_{i\in X} Z_i$. We apply matrix Bernstein's inequality.

We compute the variances $\sum_i \mathbb{E}[Z_iZ_i^\top|\Pi]$ and $\sum_i \mathbb{E}[Z_i^\top Z_i|\Pi]$. We have that $\sum_i \mathbb{E}[Z_iZ_i^\top|\Pi]$ only the diagonal terms are non-zero due to independence of Bernoulli variables, and

$$\mathbb{E}[Z_iZ_i^\top|\Pi] \le \mathrm{Diag}(F_A\pi_i) \tag{64}$$

entry-wise. Thus,

$$\begin{aligned}
\|\sum_{i\in X}\mathbb{E}[Z_iZ_i^\top|\Pi]\| &\le \max_{a\in[k]} \sum_{i\in X, b\in[k]} F_A(a,b)\pi_i(b)\\
&= \max_{a\in[k]} \sum_{i\in X, b\in[k]} F_A(a,b)\Pi_X(b,i)\\
&\le \max_{c\in[k]} \sum_{i\in X, b\in[k]} P(c,b)\Pi_X(b,d)\\
&= \|P\Pi_X\|_\infty = \|F_X\|_1.
\end{aligned} \tag{65}$$

Similarly $\sum_{i\in X}\mathbb{E}[Z_i^\top Z_i] = \sum_{i\in X}\mathrm{Diag}(\mathbb{E}[\|G_{i,A}^\top - F_A\pi_i\|^2]) \le \|F_X\|_1$. From Lemma C.12, we have $\|F_X\|_1 = O(|X|\cdot(\max_i(P\widehat{\alpha})_i))$ when $|X|$ satisfies (54).

We now bound $\|Z_i\|$. First note that the entries in $G_{i,A}$ are independent and we can use the vector Bernstein's inequality to bound $\|G_{i,A} - F_A\pi_i\|$. We have $\max_{j\in A}|G_{i,j} - (F_A\pi_i)_j| \le 2$ and $\sum_j \mathbb{E}[G_{i,j} - (F_A\pi_i)_j]^2 \le \sum_j (F_A\pi_i)_j \le \|F_A\|_1$. Thus with probability $1-\delta$, we have

$$\|G_{i,A} - F_A\pi_i\| \le (1+\sqrt{8\log(1/\delta)})\sqrt{\|F_A\|_1} + 8/3\log(1/\delta).$$

Thus, we have the bound that $\|\sum_i Z_i\| = O(\max(\sqrt{\|F_A\|_1}, \sqrt{\|F_X\|_1}))$. The concentration of the mean term follows from this result. □

We now provide spectral bounds on $\mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi]$. Define

$$\psi_i := \mathrm{Diag}(\hat{\alpha})^{-1/2}(\sqrt{\alpha_0+1}\pi_i - (\sqrt{\alpha_0+1}-1)\mu). \tag{66}$$

Let $\Psi_X$ be the matrix with columns $\psi_i$, for $i \in X$. We have

$$\mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi] = F_A\,\mathrm{Diag}(\hat{\alpha})^{1/2}\Psi_X,$$

from definition of $\mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi]$.

**Lemma C.8** (Spectral bounds)**.** *With probability $1-\delta$,*

$$\varepsilon_1 := \|I - |X|^{-1}\Psi_X\Psi_X^\top\| \le O\left(\sqrt{\frac{(\alpha_0+1)}{\widehat{\alpha}_{\min}|X|}\cdot\log\frac{k}{\delta}}\right) \tag{67}$$

*With probability* $1-2\delta$,

$$\|\mathbb{E}[(G^{\alpha_0}_{X,A})^\top|\Pi]\| = O\left(\|P\|\widehat{\alpha}_{\max}\sqrt{|X||A|(1+\varepsilon_1+\varepsilon_2)}\right)$$

$$\sigma_{\min}\left(\mathbb{E}[(G^{\alpha_0}_{X,A})^\top|\Pi]\right) = \Omega\left(\widehat{\alpha}_{\min}\sqrt{\frac{|A||X|}{\alpha_0+1}(1-\varepsilon_1-\varepsilon_3)}\cdot\sigma_{\min}(P)\cdot\right),$$

*where*

$$\varepsilon_2 := O\left(\left(\frac{1}{|A|\widehat{\alpha}^2_{\max}}\log\frac{k}{\delta}\right)^{1/4}\right), \quad \varepsilon_3 := O\left(\left(\frac{(\alpha_0+1)^2}{|A|\widehat{\alpha}^2_{\min}}\log\frac{k}{\delta}\right)^{1/4}\right). \tag{68}$$

**Remark:** When partitions $X, A$ satisfy (54), $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are small.

*Proof:* Note that $\psi_i$ is a random vector with norm bounded by $O(\sqrt{(\alpha_0+1)/\widehat{\alpha}_{min}})$ from Lemma C.12 and $\mathbb{E}[\psi_i\psi_i^\top] = I$. We now prove (67). using Matrix Bernstein Inequality. Each matrix $\psi_i\psi_i^\top/|X|$ has spectral norm at most $O((\alpha_0+1)/\widehat{\alpha}_{min}|X|)$. The variance $\sigma^2$ is bounded by

$$\left\|\frac{1}{|X|^2}\mathbb{E}[\sum_{i\in X}\|\psi_i\|^2\,\psi_i\psi_i^\top]\right\| \le \left\|\frac{1}{|X|^2}\max\|\psi_i\|^2\,\mathbb{E}[\sum_{i\in X}\psi_i\psi_i^\top]\right\| \le O((\alpha_0+1)/\widehat{\alpha}_{\min}|X|).$$

Since $O((\alpha_0+1)/\alpha_{\min}|X|) < 1$, the variance dominates in Matrix Bernstein's inequality.

Let $B := |X|^{-1}\Psi_X\Psi_X^\top$. We have with probability $1-\delta$,

$$\begin{aligned}\sigma_{\min}(\mathbb{E}[(G^{\alpha_0}_{X,A})^\top|\Pi]) &= \sqrt{|X|\sigma_{\min}(F_A\,\mathrm{Diag}(\hat{\alpha})^{1/2}B\,\mathrm{Diag}(\hat{\alpha})^{1/2}F_A^\top)},\\ &= \Omega(\sqrt{\widehat{\alpha}_{\min}|X|(1-\epsilon_1)}\cdot\sigma_{\min}(F_A)).\end{aligned}$$

From Lemma C.12, with probability $1-\delta$,

$$\sigma_{\min}(F_A) \ge \left(\sqrt{\frac{|A|\widehat{\alpha}_{\min}}{\alpha_0+1}} - O((|A|\log k/\delta)^{1/4})\right)\cdot\sigma_{\min}(P).$$

Similarly other results follow. □

## C.5 Properties of Dirichlet Distribution

In this section, we list various properties of Dirichlet distribution.

### C.5.1 Sparsity Inducing Property

We first note that the Dirichlet distribution $\mathrm{Dir}(\alpha)$ is sparse depending on values of $\alpha_i$, which is shown in [39].

**Lemma C.9.** *Let reals* $\tau\in(0,1]$, $\alpha_i>0$, $\alpha_0 := \sum_i\alpha_i$ *and integers* $1\le s\le k$ *be given. Let* $(X_i,\dots,X_k)\sim \mathrm{Dir}(\alpha)$. *Then*

$$\Pr\left[|\{i : X_i\ge\tau\}|\le s\right] \ge 1-\tau^{-\alpha_0}e^{-(s+1)/3} - e^{-4(s+1)/9},$$

*when* $s+1<3k$.

We now show that we obtain good initialization vectors under Dirichlet distribution.

Arrange the $\widehat{\alpha}_j$'s in ascending order, i.e. $\widehat{\alpha}_1 = \widehat{\alpha}_{\min} \leq \widehat{\alpha}_2 \ldots \leq \widehat{\alpha}_k = \widehat{\alpha}_{\max}$. Recall that columns vectors $\hat{W}_A^\top G_{i,A}^\top$, for $i \notin A$, are used as initialization vectors to the tensor power method. We say that $u_i := \frac{\hat{W}_A^\top G_{i,A}^\top}{\|\hat{W}_A^\top G_{i,A}^\top\|}$ is a $(\gamma, R_0)$-good initialization vector corresponding to $j \in [k]$ if

$$|\langle u_i, \Phi_j \rangle| \geq R_0, \quad |\langle u_i, \Phi_j \rangle| - \max_{m<j} |\langle u_i, \Phi_m \rangle| \geq \gamma \left|\langle u_i, \Phi_j \rangle\right|, \tag{69}$$

where $\Phi_j := \widehat{\alpha}_j^{1/2}(\tilde{F}_A)_j$, where $(\tilde{F}_A)_j$ is the $j^{\text{th}}$ column of $\tilde{F}_A := W_A^\top F_A$. Note that the $\{\Phi_j\}$ are orthonormal and are the eigenvectors to be estimated by the tensor power method.

**Lemma C.10** (Good initialization vectors under Dirichlet distribution)**.** *When $\pi_i \overset{iid}{\sim} \mathrm{Dir}(\alpha)$, and $\alpha_j < 1$, let*

$$\Delta := O\left(\frac{\zeta\rho}{\sqrt{n}r_0}\right). \tag{70}$$

*For $j \in [k]$, there is at least one $(\gamma - \frac{2\Delta}{r_0 - \Delta}, r_0 - \Delta)$-good vector corresponding to each $\Phi_j$, for $j \in [k]$, among $\{u_i\}_{i \in [n]}$ with probability $1 - 9\delta$, when*

$$n = \tilde{\Omega}\left(\alpha_{\min}^{-1} e^{r_0 \widehat{\alpha}_{\max}^{1/2}(\alpha_0 + c_1\sqrt{k\alpha_0})} (2k)^{r_0 c_2} \log(k/\delta)\right), \tag{71}$$

*where $c_1 := (1 + \sqrt{8\log 4})$ and $c_2 := 4/3(\log 4)$, when*

$$(1-\gamma) r_0 \widehat{\alpha}_{\min}^{1/2}(\alpha_0 + (1 + \sqrt{8\log 4})\sqrt{k\alpha_0} + 4/3(\log 4)\widehat{\alpha}_{\min}^{-1/2}\log 2k) > 1. \tag{72}$$

*When $\alpha_0 < 1$, the bound can be improved for $r_0 \in (0.5, (\alpha_0+1)^{-1})$ and $1 - \gamma \geq \frac{1-r_0}{r_0}$ as*

$$n > \frac{(1+\alpha_0)(1 - r_0\widehat{\alpha}_{\min})}{\widehat{\alpha}_{\min}(\alpha_{\min} + 1 - r_0(\alpha_0+1))} \log(k/\delta). \tag{73}$$

**Remark when $\alpha_0 \geq 1$, $\alpha_0 = \Theta(1)$:** When $r_0$ is chosen as $r_0 = \alpha_{\max}^{-1/2}(\sqrt{\alpha_0} + c_1\sqrt{k})^{-1}$, the term $e^{r_0 \widehat{\alpha}_{\max}^{1/2}(\alpha_0 + c_1\sqrt{k\alpha_0})} = e$, and we require

$$n = \tilde{\Omega}\left(\alpha_{\min}^{-1} k^{0.43}\log(k/\delta)\right), \quad r_0 = \alpha_{\max}^{-1/2}(\sqrt{\alpha_0} + c_1\sqrt{k})^{-1}, \tag{74}$$

by substituting $c_2/c_1 = 0.43$. Moreover, (72) is satisfied for the above choice of $r_0$ when $\gamma = \Theta(1)$.

In this case we also need $\Delta < r_0/2$, which implies

$$\zeta = O\left(\frac{\sqrt{n}}{\rho k \widehat{\alpha}_{max}}\right) \tag{75}$$

**Remark when $\alpha_0 < 1$:** In this regime, (73) implies that we require $n = \Omega(\widehat{\alpha}_{\min}^{-1})$. Also, $r_0$ is a constant, we just need $\zeta = O(\sqrt{n}/\rho)$.

*Proof:* Define $\tilde{u}_i := W_A^\top F_A \pi_i / \|W_A^\top F_A \pi_i\|$, when whitening matrix $W_A$ and $F_A$ corresponding to exact statistics are input.

We first observe that if $\tilde{u}_i$ is $(\gamma, r_0)$ good, then $u_i$ is $(\gamma - \frac{2\Delta}{r_0 - \Delta}, r_0 - \Delta)$ good.

When $\tilde{u}_i$ is $(\gamma, r_0)$ good, note that $W_A^\top F_A \pi_i \geq \widehat{\alpha}_{max}^{-1/2} r_0$ because $\sigma_{min}(W_A^\top F_A) = \widehat{\alpha}_{max}^{-1/2}$ and $\|\pi_i\| \geq r_0$. Now with probability 1/4, we have

$$\begin{aligned}
&\max_{j \in [k]} \|u_i - \tilde{u}_i\| \\
&= \left\|\hat{W}_A^\top G_{i,A} - \hat{W}_A F_A \pi_i\right\| / \left\|W_A^\top F_A \pi_i\right\| \\
&= \tilde{O}\left(\frac{\widehat{\alpha}_{\max}^{0.5}\sqrt{(\alpha_0+1)(\max_i(P\widehat{\alpha})_i)}}{r_0 n^{1/2}\widehat{\alpha}_{\min}^{1.5}\sigma_{\min}(P)}\right) = \Delta
\end{aligned}$$

from Lemma C.4. Notice that when communities are uniform $\Delta$ is proportional to $\varepsilon_{W_A}$.

If we perturb a $(\gamma, r_0)$ good vector by $\Delta$ (while maintaining unit norm), then it is still $(\gamma - \frac{2\Delta}{r_0-\Delta}, r_0 - \Delta)$ good.

We now show that the set $\{\tilde{u}_i\}$ contains good initialization vectors when $n$ is large enough. Consider $Y_i \sim \Gamma(\alpha_i, 1)$, where $\Gamma(\cdot,\cdot)$ denotes the Gamma distribution and we have $Y/\sum_i Y_i \sim \mathrm{Dir}(\alpha)$. We first compute the probability that $\tilde{u}_i := W_A^\top F_A \pi_i / \|W_A^\top F_A \pi_i\|$ is a $(r_0, \gamma)$-good vector with respect to $j = 1$ (recall that $\widehat{\alpha}_1 = \widehat{\alpha}_{\min}$). The desired event is

$$\mathcal{A}_1 := (\widehat{\alpha}_1^{-1/2} Y_1 \geq r_0 \sqrt{\sum_j \widehat{\alpha}_j^{-1} Y_j^2}) \cap (\widehat{\alpha}_1^{-1/2} Y_1 \geq \frac{1}{1-\gamma} \max_{j>1} \widehat{\alpha}_j^{-1/2} Y_j) \tag{76}$$

We have

$$\begin{aligned}
\mathbb{P}[\mathcal{A}_1] &\geq \mathbb{P}\left[(\widehat{\alpha}_{\min}^{-1/2} Y_1 \geq r_0 \sqrt{\sum_j \widehat{\alpha}_j^{-1} Y_j^2}) \cap (Y_1 \geq \frac{1}{1-\gamma} \max_{j>1} Y_j)\right] \\
&\geq \mathbb{P}\left[(\widehat{\alpha}_{\min}^{-1/2} Y_1 > r_0 t) \bigcap (\sum_j \widehat{\alpha}_j^{-1} Y_j^2 \leq t^2) \bigcap_{j>1} (Y_1 \leq (1-\gamma) r_0 t \widehat{\alpha}_{\min}^{1/2})\right], \quad \text{for some } t \\
&\geq \mathbb{P}\left[\widehat{\alpha}_{\min}^{-1/2} Y_1 > r_0 t\right] \mathbb{P}\left[\sum_j \widehat{\alpha}_j^{-1} Y_j^2 \leq t^2 \middle| \widehat{\alpha}_j^{-1/2} Y_j \leq (1-\gamma) r_0 t \widehat{\alpha}_{\min}^{1/2}\right] \mathbb{P}\left[\max_{j>1} Y_j \leq (1-\gamma) r_0 t \widehat{\alpha}_{\min}^{1/2}\right] \\
&\geq \mathbb{P}\left[\widehat{\alpha}_{\min}^{-1/2} Y_1 > r_0 t\right] \mathbb{P}\left[\sum_j \widehat{\alpha}_j^{-1} Y_j^2 \leq t^2\right] \mathbb{P}\left[\max_{j>1} Y_j \leq (1-\gamma) r_0 t \widehat{\alpha}_{\min}^{1/2}\right]
\end{aligned}$$

When $\alpha_j \leq 1$, we have

$$\mathbb{P}[\cup_j Y_j \geq \log 2k] \leq 0.5,$$

since $P(Y_j \geq t) \leq t^{\alpha_j - 1} e^{-t} \leq e^{-t}$ when $t > 1$ and $\alpha_j \leq 1$. Applying vector Bernstein's inequality, we have with probability $0.5 - e^{-m}$ that

$$\|\operatorname{Diag}(\widehat{\alpha}_j^{-1/2})(Y - \mathbb{E}(Y))\|_2 \leq (1 + \sqrt{8m})\sqrt{k\alpha_0} + 4/3 m \widehat{\alpha}_{\min}^{-1/2} \log 2k,$$

since $\mathbb{E}[\sum_j \widehat{\alpha}_j^{-1} \operatorname{Var}(Y_j)] = k\alpha_0$ since $\widehat{\alpha}_j = \alpha_j/\alpha_0$ and $\operatorname{Var}(Y_j) = \alpha_j$. Thus, we have

$$\|\operatorname{Diag}(\widehat{\alpha}_j^{-1/2}) Y\|_2 \leq \alpha_0 + (1 + \sqrt{8m})\sqrt{k\alpha_0} + 4/3 m \widehat{\alpha}_{\min}^{-1/2} \log 2k,$$

since $\|\operatorname{Diag}(\widehat{\alpha}_j^{-1/2})\mathbb{E}(Y)\|_2 = \sqrt{\sum_j \widehat{\alpha}_j^{-1} \alpha_j^2} = \alpha_0$. Choosing $m = \log 4$, we have with probability $1/4$ that

$$\begin{aligned}
\|\operatorname{Diag}(\widehat{\alpha}_j^{-1/2}) Y\|_2 \leq t &:= \alpha_0 + (1 + \sqrt{8 \log 4})\sqrt{k\alpha_0} + 4/3(\log 4) \widehat{\alpha}_{\min}^{-1/2} \log 2k, \qquad (77) \\
&= \alpha_0 + c_1 \sqrt{k\alpha_0} + c_2 \widehat{\alpha}_{\min}^{-1/2} \log 2k. \qquad (78)
\end{aligned}$$

We now have

$$\mathbb{P}\left[\widehat{\alpha}_{\min}^{-1/2} Y_1 > r_0 t\right] \geq \frac{\alpha_{\min}}{4C} \left(r_0 t \widehat{\alpha}_{\min}^{1/2}\right)^{\alpha_{\min} - 1} e^{-r_0 t \widehat{\alpha}_{\min}^{1/2}},$$

from Lemma C.13.

Similarly,

$$\mathbb{P}\left[\max_{j \neq 1} Y_j \leq \widehat{\alpha}_{\min}^{1/2}(1-\gamma) r_0 t\right] \geq 1 - \sum_j \left((1-\gamma) r_0 t \widehat{\alpha}_{\min}^{1/2}\right)^{\sum_j \alpha_j - 1} e^{-(1-\gamma) r_0 \widehat{\alpha}_{\min}^{1/2} t} \geq 1 - k e^{-(1-\gamma) r_0 \widehat{\alpha}_{\min}^{1/2} t},$$

assuming that $(1-\gamma)r_0\widehat{\alpha}_{\min}^{1/2}t > 1$.

Choosing $t$ as in (77), we have the probability of the event in (76) is greater than

$$\frac{\alpha_{\min}}{16C}\left(1-\frac{e^{-(1-\gamma)r_0\widehat{\alpha}_{\min}^{1/2}(\alpha_0+c_1\sqrt{k\alpha_0})}}{2(2k)^{(1-\gamma)r_0c_2-1}}\right)\frac{e^{-r_0\widehat{\alpha}_{\min}^{1/2}(\alpha_0+c_1\sqrt{k\alpha_0})}}{(2k)^{r_0c_2}}\left(r_0\widehat{\alpha}_{\min}^{1/2}(\alpha_0+c_1\sqrt{k\alpha_0}+c_2\widehat{\alpha}_{\min}^{-1/2}\log 2k)\right)^{\alpha_{\min}-1}$$

Similarly the (marginal) probability of events $\mathcal{A}_2$ can be bounded from below by replacing $\alpha_{\min}$ with $\alpha_2$ and so on. Thus, we have

$$\mathbb{P}[\mathcal{A}_m] = \tilde{\Omega}\left(\alpha_{\min}\frac{e^{-r_0\widehat{\alpha}_{\max}^{1/2}(\alpha_0+c_1\sqrt{k\alpha_0})}}{(2k)^{r_0c_2}}\right),$$

for all $m \in [k]$.

Thus, we have each of the events $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$ occur at least once in $n$ i.i.d. tries with probability

$$\begin{aligned}
&1-\mathbb{P}\left[\bigcup_{j\in[k]}\Big(\bigcap_{i\in[n]}\mathcal{A}_j^c(i)\Big)\right]\\
&\geq 1-\sum_{j\in[k]}\mathbb{P}\left[\bigcap_{i\in[n]}\mathcal{A}_j^c(i)\right]\\
&\geq 1-\sum_{j\in[k]}\exp\left[-n\mathbb{P}(\mathcal{A}_j)\right],\\
&\geq 1-k\exp\left[n\tilde{\Omega}\left(\alpha_{\min}\frac{e^{-r_0\widehat{\alpha}_{\max}^{1/2}(\alpha_0+c_1\sqrt{k\alpha_0})}}{(2k)^{r_0c_2}}\right)\right]
\end{aligned}$$

where $\mathcal{A}_j(i)$ denotes the event that $\mathcal{A}_1$ occurs for $i^{\text{th}}$ trial and we use that $1-x \leq e^{-x}$ when $x \in [0,1]$. Thus, for the event to occur with probability $1-\delta$, we require

$$n = \tilde{\Omega}\left(\alpha_{\min}^{-1}e^{r_0\widehat{\alpha}_{\max}^{1/2}(\alpha_0+c_1\sqrt{k\alpha_0})}(2k)^{r_0c_2}\log(1/\delta)\right).$$

**Improved Bound when $\alpha_0 < 1$:** We can improve the above bound by directly working with the Dirichlet distribution. Let $\pi \sim \mathrm{Dir}(\alpha)$. The desired event corresponding to $j=1$ is given by

$$\mathcal{A}_1 = \left(\frac{\widehat{\alpha}_1^{-1/2}\pi_1}{\|\operatorname{Diag}(\widehat{\alpha}_i^{-1/2})\pi\|} \geq r_0\right)\bigcap_{i>1}\left(\pi_1 \geq \frac{\pi_i}{1-\gamma}\right).$$

Thus, we have

$$\begin{aligned}
\mathbb{P}[\mathcal{A}_1] &\geq \mathbb{P}\left[(\pi_1 \geq r_0)\bigcap_{i>1}(\pi_i \leq (1-\gamma)r_0)\right]\\
&\geq \mathbb{P}[\pi_1 \geq r_0]\mathbb{P}\left(\bigcap_{i>1}\pi_i \leq (1-\gamma)r_0|\pi_1 \geq r_0\right),
\end{aligned}$$

since $\mathbb{P}\left(\bigcap_{i>1}\pi_i \leq (1-\gamma)r_0|\pi_1 \geq r_0\right) \geq \mathbb{P}\left(\bigcap_{i>1}\pi_i \leq (1-\gamma)r_0\right)$. By properties of Dirichlet distribution, we know $\mathbb{E}[\pi_i] = \widehat{\alpha}_i$ and $\mathbb{E}[\pi_i^2] = \widehat{\alpha}_i\frac{\alpha_i+1}{\alpha_0+1}$. Let $p := \Pr[\pi_1 \geq r_0]$. We have

$$\begin{aligned}
\mathbb{E}[\pi_i^2] &= p\mathbb{E}[\pi_i^2|\pi_i \geq r_0] + (1-p)\mathbb{E}[\pi_i^2|\pi_i < r_0]\\
&\leq p + (1-p)r_0\mathbb{E}[\pi_i|\pi_i < r_0]\\
&\leq p + (1-p)r_0\mathbb{E}[\pi_i]
\end{aligned}$$

Thus, $p \geq \frac{\widehat{\alpha}_{\min}(\alpha_{\min}+1-r_0(\alpha_0+1))}{(\alpha_0+1)(1-r_0\widehat{\alpha}_{\min})}$, which is useful when $r_0(\alpha_0+1) < 1$. Also when $\pi_1 \geq r_0$, we have that $\pi_i \leq 1 - r_0$ since $\pi_i \geq 0$ and $\sum_i \pi_i = 1$. Thus, choosing $1-\gamma = \frac{1-r_0}{r_0}$, we have the other conditions for $\mathcal{A}_1$ are satisfied. Also, verify that we have $\gamma < 1$ when $r_0 > 0.5$ and this is feasible when $\alpha_0 < 1$. $\square$

We now prove a result that the entries of $\pi_i$, which are marginals of the Dirichlet distribution, are likely to be small in the sparse regime of the Dirichlet parameters. Recall that the marginal distribution of $\pi_i$ is distributed as $B(\alpha_i, \alpha_0 - \alpha_i)$, where $B(a, b)$ is the beta distribution and

$$\mathbb{P}[Z = z] \propto z^{a-1}(1-z)^{b-1}, \quad Z \sim B(a, b).$$

**Lemma C.11** (Marginal Dirichlet distribution in sparse regime)**.** *For $Z \sim B(a, b)$, the following results hold:*

**Case $b \leq 1$, $C \in [0, 1/2]$:**

$$\Pr[Z \geq C] \leq 8 \log(1/C) \cdot \frac{a}{a+b} \tag{79}$$

$$\mathbb{E}[Z \cdot \delta(Z \leq C)] \leq C \cdot \mathbb{E}[Z] = C \cdot \frac{a}{a+b} \tag{80}$$

**Case $b \geq 1$, $C \leq (b+1)^{-1}$:** *we have*

$$\Pr[Z \geq C] \leq a \log(1/C) \tag{81}$$

$$\mathbb{E}[Z \cdot \delta(Z \leq C)] \leq 6aC \tag{82}$$

**Remark:** The guarantee for $b \geq 1$ is worse and this agrees with the intuition that the Dirichlet vectors are more spread out (or less sparse) when $b = \alpha_0 - \alpha_i$ is large.

*Proof.* We have

$$\begin{aligned}
\mathbb{E}[Z \cdot \delta(Z \leq C)] &= \int_0^C \frac{1}{B(a,b)} x^a (1-x)^{b-1} dx \\
&\leq \frac{(1-C)^{b-1}}{B(a,b)} \int_0^C x^a dx \\
&= \frac{(1-C)^{b-1} C^{a+1}}{(a+1)B(a,b)}
\end{aligned}$$

For $\mathbb{E}[Z \cdot \delta(Z \geq C)]$, we have,

$$\begin{aligned}
\mathbb{E}[Z \cdot \delta(Z \geq C)] &= \int_C^1 \frac{1}{B(a,b)} x^a (1-x)^{b-1} dx \\
&\geq \frac{C^a}{B(a,b)} \int_C^1 (1-x)^{b-1} dx \\
&= \frac{(1-C)^b C^a}{bB(a,b)}
\end{aligned}$$

The ratio between these two is at least

$$\frac{\mathbb{E}[Z \cdot \delta(Z \geq C)]}{\mathbb{E}[Z \cdot \delta(Z \leq C)]} \geq \frac{(1-C)(a+1)}{bC} \geq \frac{1}{C}.$$

The last inequality holds when $a, b < 1$ and $C < 1/2$. The sum of the two is exactly $\mathbb{E}[Z]$, so when $C < 1/2$ we know $\mathbb{E}[Z \cdot \delta(Z \leq C)] < C \cdot \mathbb{E}[Z]$.

Next we bound the probability $\Pr[Z \geq C]$. Note that $\Pr[Z \geq 1/2] \leq 2\mathbb{E}[Z] = \frac{2a}{a+b}$ by Markov's inequality. Now we show $\Pr[Z \in [C, 1/2]]$ is not much larger than $\Pr[Z \geq 1/2]$ by bounding the integrals.

$$A = \int_{1/2}^{1} x^{a-1}(1-x)^{b-1}dx \geq \int_{1/2}^{1} (1-x)^{b-1}dx = (1/2)^b/b.$$

$$\begin{aligned} B = \int_{C}^{1/2} x^{a-1}(1-x)^{b-1} &\leq (1/2)^{b-1} \int_{C}^{1/2} x^{a-1}dx \\ &\leq (1/2)^{b-1}\frac{0.5^a - C^a}{a} \\ &\leq (1/2)^{b-1}\frac{1-(1-a\log 1/C)}{a} \\ &= (1/2)^{b-1}\log(1/C). \end{aligned}$$

The last inequality uses the fact that $e^x \geq 1+x$ for all $x$. Now

$$\Pr[Z \geq C] = (1+\frac{B}{A})\Pr[Z \geq 1/2] \leq (1+2b\log(1/C))\frac{2a}{a+b} \leq 8\log(1/C)\cdot\frac{a}{a+b}$$

and we have the result.

**Case 2:** When $b \geq 1$, we have an alternative bound. We use the fact that if $X \sim \Gamma(a,1)$ and $Y \sim \Gamma(b,1)$ then $Z \sim X/(X+Y)$. Since $Y$ is distributed as $\Gamma(b,1)$, its PDF is $\frac{1}{\Gamma(b)}x^{b-1}e^{-x}$. This is proportional to the PDF of $\Gamma(1)$ ($e^{-x}$) multiplied by a increasing function $x^{b-1}$.

Therefore we know $\Pr[Y \geq t] \geq \Pr_{Y' \sim \Gamma(1)}[Y' \geq t] = e^{-t}$.

Now we use this bound to compute the probability that $Z \leq 1/R$ for all $R \geq 1$.

This is equivalent to

$$\begin{aligned} \Pr[\frac{X}{X+Y} \leq \frac{1}{R}] &= \int_0^\infty Pr[X=x]Pr[Y \geq (R-1)X]dx \\ &\geq \int_0^\infty \frac{1}{\Gamma(a)}x^{a-1}e^{-Rx}dx \\ &= R^{-a}\int_0^\infty \frac{1}{\Gamma(a)}y^{a-1}e^{-y}dy \\ &= R^{-a} \end{aligned}$$

In particular, $\Pr[Z \leq C] \geq C^a$, which means $\Pr[Z \geq C] \leq 1 - C^a \leq a\log(1/C)$.

For $\mathbb{E}[Z\delta(Z < C)]$, the proof is similar as before:

$$P = \mathbb{E}[Z\delta(Z < C)] = \int_0^C \frac{1}{B(a,b)}x^a(1-x)^b dx \leq \frac{C^{a+1}}{B(a,b)(a+1)}$$

$$Q = \mathbb{E}[Z\delta(Z \geq C)] = \int_C^1 \frac{1}{B(a,b)}x^a(1-x)^b dx \geq \frac{C^a(1-C)^{b+1}}{B(a,b)(b+1)}$$

Now $\mathbb{E}[Z\delta(Z \leq C)] \leq \frac{P}{Q}\mathbb{E}[Z] \leq 6aC$ when $C < 1/(b+1)$. □

### C.5.2 Norm Bounds

**Lemma C.12** (Norm Bounds under Dirichlet distribution)**.** *For $\pi_i \overset{iid}{\sim} \mathrm{Dir}(\alpha)$ for $i \in A$, with probability $1-\delta$, we have*

$$\sigma_{\min}(\Pi_A) \geq \sqrt{\frac{|A|\widehat{\alpha}_{\min}}{\alpha_0+1}} - O((|A|\log k/\delta)^{1/4}),$$
$$\|\Pi_A\| \leq \sqrt{|A|\widehat{\alpha}_{\max}} + O((|A|\log k/\delta)^{1/4}),$$
$$\kappa(\Pi_A) \leq \sqrt{\frac{(\alpha_0+1)\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}} + O((|A|\log k/\delta)^{1/4}).$$

*This implies that $\|F_A\| \leq \|P\| \sqrt{|A|\widehat{\alpha}_{\max}}$, $\kappa(F_A) \leq O(\kappa(P)\sqrt{(\alpha_0+1)\widehat{\alpha}_{\max}/\widehat{\alpha}_{min}})$. Moreover, with probability $1-\delta$*

$$\|F_A\|_1 \leq |A| \cdot \max_i (P\widehat{\alpha})_i + O\left(\|P\|\sqrt{|A|\log\frac{|A|}{\delta}}\right) \tag{83}$$

**Remark:** When $|A| = \Omega\left(\log\frac{k}{\delta}\left(\frac{\alpha_0+1}{\widehat{\alpha}_{\min}}\right)^2\right)$, we have $\sigma_{\min}(\Pi_A) = \Omega(\sqrt{\frac{|A|\widehat{\alpha}_{\min}}{\alpha_0+1}})$ with probability $1-\delta$ for any fixed $\delta \in (0,1)$.

*Proof:* Consider $\Pi_A \Pi_A^\top = \sum_{i\in A} \pi_i \pi_i^\top$.

$$\begin{aligned}\frac{1}{|A|}\mathbb{E}[\Pi_A^\top \Pi_A] =& \mathbb{E}_{\pi\sim Dir(\alpha)}[\pi\pi^\top] \\ =& \frac{\alpha_0}{\alpha_0+1}\widehat{\alpha}\widehat{\alpha}^\top + \frac{1}{\alpha_0+1}\operatorname{Diag}(\widehat{\alpha}),\end{aligned}$$

from Proposition C.14. The first term is positive semi-definite so the eigenvalues of the sum are at least the eigenvalues of the second component. Smallest eigenvalue of second component gives lower bound on $\sigma_{\min}(\mathbb{E}[\Pi_A\Pi_A^\top])$. The spectral norm of the first component is bounded by $\frac{\alpha_0}{\alpha_0+1}\|\hat{\alpha}\| \leq \frac{\alpha_0}{\alpha_0+1}\widehat{\alpha}_{\max}$, the spectral norm of second component is $\frac{1}{\alpha_0+1}\alpha_{\max}$. Thus $\left\|\mathbb{E}[\Pi_A\Pi_A^\top]\right\| \leq |A| \cdot \widehat{\alpha}_{\max}$.

Now applying Matrix Bernstein's inequality to $\frac{1}{|A|}\sum_i \left(\pi_i\pi_i^\top - \mathbb{E}[\pi\pi^\top]\right)$. We have that the variance is $O(1/|A|)$. Thus with probability $1-\delta$,

$$\left\|\frac{1}{|A|}\left(\Pi_A\Pi_A^\top - \mathbb{E}[\Pi_A\Pi_A^\top]\right)\right\| = O\left(\sqrt{\frac{\log(k/\delta)}{|A|}}\right).$$

For the result on $F$, we use the property that for any two matrices $A, B$, $\|AB\| \leq \|A\|\,\|B\|$ and $\kappa(AB) \leq \kappa(A)\kappa(B)$.

To show bound on $\|F_A\|_1$, note that each column of $F_A$ satisfies $\mathbb{E}[(F_A)_i] = \langle\widehat{\alpha}, (P)_i\rangle\, 1^\top$, and thus $\|\mathbb{E}[F_A]\|_1 \leq |A| \max_i (P\widehat{\alpha})_i$. Using Bernstein's inequality, for each column of $F_A$, we have, with probability $1-\delta$,

$$|\,\|(F_A)_i\|_1 - |A|\,\langle\widehat{\alpha}, (P)_i\rangle| = O\left(\|P\|\sqrt{|A|\log\frac{|A|}{\delta}}\right),$$

by applying Bernstein's inequality, since $|\,\langle\widehat{\alpha}, (P)_i\rangle\,| \leq \|P\|$, and thus we have $\sum_{i\in A}\|\mathbb{E}[(P)_j^\top \pi_i \pi^\top (P)_j]\|$, and $\sum_{i\in A}\|\mathbb{E}[\pi_i^\top (P)_j (P)_j^\top \pi]\| \leq |A| \cdot \|P\|$. $\square$

### C.5.3 Properties of Gamma and Dirichlet Distributions

Recall Gamma distribution $\Gamma(\alpha, \beta)$ is a distribution on nonnegative real values with density function $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.

**Proposition C.13** (Dirichlet and Gamma distributions)**.** *The following facts are known for Dirichlet distribution and Gamma distribution.*

1. *Let $Y_i \sim \Gamma(\alpha_i, 1)$ be independent random variables, then the vector $(Y_1, Y_2, ..., Y_k)/\sum_{i=1}^k Y_k$ is distributed as $Dir(\alpha)$.*
2. *The $\Gamma$ function satisfies Euler's reflection formula: $\Gamma(1-z)\Gamma(z) \le \pi/\sin \pi_z$.*
3. *The $\Gamma(z) \ge 1$ when $0 < z < 1$.*
4. *There exists a universal constant $C$ such that $\Gamma(z) \le C/z$ when $0 < z < 1$.*
5. *For $Y \sim \Gamma(\alpha, 1)$ and $t > 0$ and $\alpha \in (0, 1)$, we have*
$$\frac{\alpha}{4C} t^{\alpha-1} e^{-t} \le \Pr[Y \ge t] \le t^{\alpha-1} e^{-t}, \tag{84}$$
*and for any $\eta, c > 1$, we have*
$$\mathbb{P}[Y > \eta t | Y \ge t] \ge (c\eta)^{\alpha-1} e^{-(\eta-1)t}. \tag{85}$$

*Proof:* The bounds in (84) is derived using the fact that $1 \le \Gamma(\alpha) \le C/\alpha$ when $\alpha \in (0, 1)$ and
$$\int_t^\infty \frac{1}{\Gamma(\alpha_i)} x^{\alpha_i-1} e^{-x} dx \le \frac{1}{\Gamma(\alpha_i)} \int_t^\infty t^{\alpha_i-1} e^{-x} dx \le t^{\alpha_i-1} e^{-t},$$
and
$$\int_t^\infty \frac{1}{\Gamma(\alpha_i)} x^{\alpha_i-1} e^{-x} dx \ge \frac{1}{\Gamma(\alpha_i)} \int_t^{2t} x^{\alpha_i-1} e^{-x} dx \ge \alpha_i/C \int_t^{2t} (2t)^{\alpha_i-1} e^{-x} dx \ge \frac{\alpha_i}{4C} t^{\alpha_i-1} e^{-t}.$$
□

**Proposition C.14** (Moments under Dirichlet distribution)**.** *Suppose $v \sim Dir(\alpha)$, the moments of $v$ satisfies the following formulas:*
$$\begin{aligned} \mathbb{E}[v_i] &= \frac{\alpha_i}{\alpha_0} \\ \mathbb{E}[v_i^2] &= \frac{\alpha_i(\alpha_i+1)}{\alpha_0(\alpha_0+1)} \\ \mathbb{E}[v_i v_j] &= \frac{\alpha_i \alpha_j}{\alpha_0(\alpha_0+1)}, \quad i \ne j. \end{aligned}$$

*More generally, if $a^{(t)} = \prod_{i=0}^{t-1}(a+i)$, then we have*
$$\mathbb{E}[\prod_{i=1}^k v_i^{(a_i)}] = \frac{\prod_{i=1}^k \alpha_i^{(a_i)}}{\alpha_0^{(\sum_{i=1}^k a_i)}}.$$

## C.6 Standard Results

**Bernstein's Inequalities:** One of the key tools we use is the standard matrix Bernstein inequality [40, thm. 1.4].

**Proposition C.15** (Matrix Bernstein Inequality)**.** *Suppose* $Z = \sum_j W_j$ *where*

1. $W_j$ *are independent random matrices with dimension* $d_1 \times d_2$,
2. $\mathbb{E}[W_j] = 0$ *for all* $j$,
3. $\|W_j\| \leq R$ *almost surely.*

*Let* $d = d_1 + d_2$, *and* $\sigma^2 = \max\left\{\left\|\sum_j \mathbb{E}[W_j W_j^\top]\right\|, \left\|\sum_j \mathbb{E}[W_j^\top W_j]\right\|\right\}$, *then we have*

$$\Pr[\|Z\| \geq t] \leq d \cdot exp\left\{\frac{-t^2/2}{\sigma^2 + Rt/3}\right\}.$$

**Proposition C.16** (Vector Bernstein Inequality)**.** *Let* $z = (z_1, z_2, ..., z_n) \in \mathbb{R}^n$ *be a random vector with independent entries,* $\mathbb{E}[z_i] = 0$, $\mathbb{E}[z_i^2] = \sigma_i^2$, *and* $\Pr[|z_i| \leq 1] = 1$. *Let* $A = [a_1|a_2|\cdots|a_n] \in \mathbb{R}^{m\times n}$ *be a matrix, then*

$$\Pr[\|Az\| \leq (1+\sqrt{8t})\sqrt{\sum_{i=1}^{n} \|a_i\|^2 \sigma_i^2} + (4/3)\max_{i\in[n]} \|a_i\|\, t] \geq 1 - e^{-t}.$$

**Wedin's theorem:** We make use of Wedin's theorem to control subspace perturbations.

**Lemma C.17** (Wedin's theorem; Theorem 4.4, p. 262 in [38].)**.** *Let* $A, E \in \mathbb{R}^{m\times n}$ *with* $m \geq n$ *be given. Let* $A$ *have the singular value decomposition*

$$\begin{bmatrix} U_1^\top \\ U_2^\top \\ U_3^\top \end{bmatrix} A \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{bmatrix}.$$

*Let* $\tilde{A} := A + E$, *with analogous singular value decomposition* $(\tilde{U}_1, \tilde{U}_2, \tilde{U}_3, \tilde{\Sigma}_1, \tilde{\Sigma}_2, \tilde{V}_1\tilde{V}_2)$. *Let* $\Phi$ *be the matrix of canonical angles between* $\mathrm{range}(U_1)$ *and* $\mathrm{range}(\tilde{U}_1)$, *and* $\Theta$ *be the matrix of canonical angles between* $\mathrm{range}(V_1)$ *and* $\mathrm{range}(\tilde{V}_1)$. *If there exists* $\delta, \alpha > 0$ *such that* $\min_i \sigma_i(\tilde{\Sigma}_1) \geq \alpha + \delta$ *and* $\max_i \sigma_i(\Sigma_2) \leq \alpha$, *then*

$$\max\{\|\sin\Phi\|_2, \|\sin\Theta\|_2\} \leq \frac{\|E\|_2}{\delta}.$$